Support of Resource-Aware Vertical Handovers in WLAN Hotspots

vorgelegt von
Diplom-Ingenieur
Sven Wiethölter
geb. in Berlin

von der Fakultät IV – Elektrotechnik und Informatik –
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. Dr.-Ing. Thomas Sikora
Gutachter: Prof. Dr.-Ing. Adam Wolisz
Gutachter: Prof. Dr.-Ing. Andreas Timm-Giel
Gutachter: Prof. Dr. Axel Küpper

Tag der wissenschaftlichen Aussprache: 21. Mai 2014

Berlin 2014
D 83

Zusammenfassung

Endgeräte wie Smartphones oder Tablets bieten häufig eine Vielfalt drahtloser Zugänge zum Internet an. Üblicherweise schließt dies die 802.11 WLANs und auch Technologien drahtloser Weitverkehrsnetze (WWANs) aus dem Bereich LTE oder WiMAX ein. Aufgrund dieser Optionen haben sich die Endanwender daran gewöhnt, überall und zu jeder Zeit auf ihre Internetdienste zuzugreifen. Damit hat auch der Datenverkehr pro Anwender zugenommen, was eine Herausforderung insbesondere für die Betreiber von WWANs ist. Soweit verfügbar, favorisieren Endanwender heutzutage eher einen drahtlosen Zugang zum Internet über WLANs als über WWANs. Des Weiteren haben die 3GPP-Standardisierungsgremien Ansätze erarbeitet, die zusätzlich Verkehr aus WWANs in Netze mit geringerer Abdeckung wie WLAN- oder Femto-Zellen abgeben. Solche Ansätze werden auch als 'Traffic Offloading' bezeichnet und haben das Ziel, die WWANs zu entlasten. Dabei werden jedoch eher einfache Strategien verfolgt, die auf der Nutzung zusätzlicher Kapazitäten heterogener Netze beruhen und dann angewendet werden, wenn ein alternatives Zugangsnetz für ein Endgerät verfügbar ist. Im Rahmen dieser Arbeit zeigen wir Gewinne auf, die entstehen, wenn man die Auswahl der Endgeräte für ein WLAN-Netz stattdessen auf Basis der von ihnen belegten Ressourcen durchführt. In diesem Kontext schlagen wir vor, Geräte mit stark negativem Einfluss auf die WLAN-Kapazität wieder zurück in das WWAN zu reichen, was wir als 'Onloading' bezeichnen. Ein solches 'Onloading' zieht Herausforderungen in unterschiedlichen Richtungen mit sich.

Die fortschreitende Miniaturisierung hat in den letzten Jahren zu dem Trend geführt, die Anzahl der Netzwerkkarten (NICs) in Endgeräten zu reduzieren. Wir bezeichnen eine NIC als multimodal, wenn sie mehrere Funktechnologien unterstützt, aber zu einem bestimmten Zeitpunkt immer nur eine davon genutzt werden kann. Deswegen stellt für eine multimodale NIC das 'Onloading' während einer laufenden Verbindung eine Herausforderung dar. Wir schlagen einen Ansatz vor, der vorbereitende Mechanismen für ein 'Onloading' als auch eine laufende Verbindung im WLAN über eine solche NIC ermöglicht.

Des Weiteren ist es wichtig, in einem WLAN Hotspot zu entscheiden, welche Geräte einen negativen Einfluss auf die Kapazität des Netzes haben. Dafür haben wir eine Metrik entwickelt, die eine Entscheidungsgrundlage für das Onloading bildet. Diese Metrik basiert rein auf einer Beobachtung des Netzes und seiner Geräte, ermöglicht jedoch keine Entscheidung für sich neu assoziierende Geräte im WLAN. Erschwerend kommt hinzu, dass viele Eigenschaften der NICs durch herstellerabhängige Implementierungen geprägt werden. Solche Algorithmen bieten eine zusätzliche Herausforderung, da ihre internen Abläufe üblicherweise unbekannt sind. Ein bekanntes Beispiel für solche Algorithmen stellt die Anpassung der WLAN-Link-Datenraten dar. Diese Algorithmen wählen die jeweiligen Modulations- und Kodierungsschemata (MCSs) für die drahtlosen Übertragungen aus. Robuste MCSs resultieren dabei in geringere Link-Datenraten und haben somit einen starken Einfluss auf die Kapazität einer WLAN-Zelle. Aus diesem Grund fokussieren wir uns auf eine Abschätzung der Datenratenwahl eines Endgerätes. Damit lassen sich im Vorfeld Aussagen treffen, ob ein Gerät starken Einfluss auf die WLAN-Kapazität haben wird, so dass es für ein 'Onloading' in Frage kommt.

Abstract

End-user devices such as smart phones and tablets have become very popular as they offer a variety of wireless Internet accesses ranging from the WLAN standards to WWAN technologies such as LTE or even WiMAX. Due to these different wireless access options and new emerging applications—e.g., from the areas of video streaming, social networks, as well as Internet clouds—people are increasingly connecting to the Internet with their devices while being on the move. In line with this, the number of devices as well as the traffic demand of end users have been reported to increase rapidly over the last years which imposes a strong challenge especially for the operators of WWANs. Thereby, end users frequently tend to use settings that favor a connectivity to the Internet whenever possible rather over WLAN than over WWAN access. Further, the cellular standardization bodies of the 3GPP envision solutions to hand over on-going wireless sessions from cellular to other small cell accesses such as WLANs or femto cells. This is also known as *traffic offloading* essentially freeing capacity in terms of users with a certain service in the cellular accesses. Nevertheless this offloading follows a rather simple strategy to utilize additional capacity of heterogeneous accesses such as WLANs whenever being available for a given device. This thesis shows that stronger gains can be expected if the selection of devices to be served in WLANs is conducted in a resource-aware fashion including an evaluation of the WLAN traffic in terms of the channel occupation time and MAC overhead as result of contention, interference, and fluctuating channels. In this context, this thesis envisions to *onload* unfavorable devices negatively affecting the WLAN capacity back to WWAN accesses. A support of such an onloading imposes challenges in different dimensions.

From the hardware design of devices, there is a strong trend to limit the number of separate network interface cards (NICs) due to space and cost issues. We refer to a multi-mode NIC if it covers multiple technologies, while at a given time only access to one technology is possible. Thus, smoothly onloading a device with such a NIC is by far not trivial. We present an approach that conducts handover preparation mechanisms, while also allowing a continuous WLAN communication over a multi-mode NIC.

Further, it is by far not trivial to judge which subset of associated devices is negatively affecting the capacity of a WLAN hotspot. Thus, a careful evaluation of devices regarding a selection for an onloading back to WWAN accesses imposes a challenge yet. In this direction, we present a performance metric that identifies devices degrading the WLAN capacity. While our performance metric tackles a reactive selection, it falls short to support a predictive evaluation, e.g., of devices which just joined the WLAN cell. Even worse, proprietary algorithms inside a WLAN stack impose a severe challenge as their internal routines are usually not conveyed via typical management interfaces. A well-known example for this category of algorithms are the link data rate adaptation schemes, with which WLAN devices adjust the modulation and coding scheme (MCS) for their transmissions. As MCSs resulting in low link data rates may specifically degrade the capacity of a WLAN cell, we focus on an estimation regarding the data rate selection of a device as a third contribution of this thesis. This estimation enables to select devices that will likely degrade the capacity of the WLAN hotspot for an onloading in advance.

Acknowledgments

First and foremost, I thank Prof. Adam Wolisz for his support throughout all my years at TKN. This does not only include the financial perspective or the topical parts including all the excellent feedbacks and discussions; having a 'Doktorvater' on whom one can rely in all kinds of situations is probably the most solid foundation. Further, I take the opportunity to thank the further reviewers of my thesis, Prof. Andreas Timm-Giel and Prof. Axel Küpper. Thanks for accepting the review of my thesis and for your valuable feedback. In addition, I thank Prof. Thomas Sikora for chairing the thesis committee.

My work regarding the link-data rate estimation benefited a lot from a cooperation with Prof. Manfred Opper and Dr. Andreas Ruttor. Thanks a lot to both of you for pointing me to the opportunities of Gaussian Processes regarding my specific issue, our in-depth discussions, and all your support. It really has been fun to work with you.

At a certain stage of writing a thesis, it becomes indispensable to have people reviewing the whole pile of paper regarding content, style, and grammar. A big thanks goes to my 'review buddies' Marc Emmelmann, Dr. Łukasz Budzisz, and my sister Dr. Anke Wiethölter. Further, I specifically thank Marc Emmelmann for our joint time, in- and outside of TKN. Besides our private activities, this includes all the countless hours of discussions about our research that have led to so many joint papers.

Since my first academic days at TU Berlin, Marc Löbbers, Dr. Daniel Willkomm, and Dr. Mathias Bohge accompany my way. Having a look back today, it is just amazing what we have experienced altogether, e.g., learning sessions, (student) parties, and the joint time at TKN. I am really excited to see what's coming up next with you, guys.

At TKN, I have experienced a very productive and friendly working atmosphere. Thanks for the nice time both to all current and previous TKN colleagues, specifically to Konstantin Miller, Thomas Menzel, Dr. Filip Idzikowski, Dr. Jan Hauer, Onur Ergin, Niels Karowski, Dr. Andreas Köpke, Sven Hermann, Lars Westerhoff, Tobias Poschwatta, Andreas Köpsel, and Dr. Murad Abusubaih.

For their commitment and their work, I warmly thank 'my' students, namely Yerong Chen, Uwe Bergemann, Joseph Cherukara, Robert Andersson, Hyung-Taek Lim, Jan Scheer, and Svetoslav Yankov. I wish you all the best for your further career.

For all my simulations and testbed setups, I was able to rely on the continuous support of the technical staff team, namely Sven Spuida, Georgios Ainatzes, Peter Schröter, Jürgen Malinowski, and Berthold Rathke. Thanks for your help regarding the computer infrastructure and the lab environment.

As everywhere, organizational issues may be cumbersome and time consuming. Thanks to the administrative assistances Petra Hutt, Heike Klemz, and Sonja Cecar for helping out regarding various issues thus minimizing the administrative effort on my side.

Further, I thank my parents and my sister Anke for their continuous support throughout my whole life. Independent of the actual conditions, including both calm and stormy situations, I have been capable to always count and rely on you. Finally, I thank my wife for her motivational support, for keeping me free for my thesis project, as well as for accepting my countless hours of work on weekends and during nights. Madena, without

you, I probably would have never completed this thesis. And together with you, the year 2014 has really become one of the best ones. Thank you for all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Individual devices (e.g., laptops, smart phones, or tablets) use a diversity of heterogeneous wireless access technologies that has been growing continuously over the past few years. Technologies from the area of both wireless local area networks (WLANs) and wireless wide area networks (WWANs) count nowadays as a standard mix for these devices. From a historical perspective, WLANs and WWANs differ significantly with respect to their coverage area as well as their support regarding mobility and quality of service (QoS) for each device. WLANs usually offer a hotspot access covering small areas without any mobility support. In contrast, WWANs can even maintain a wireless connectivity at a high device mobility degree over large coverage areas, thus enabling a continuous QoS support. Today, end devices usually support wireless access for WLANs as specified by the Institute of Electrical and Electronic Engineers (IEEE) 802.11 standard, while the WWAN technology may comprise of 3rd Generation Partnership Project (3GPP) Universal Mobile Telecommunications System (UMTS) and Long Term Evolution (LTE) or even IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX) and related extensions.

The wide deployment of devices with different wireless access technologies has several implications. On the one hand, users depend increasingly on the preferably high speed connectivity for everyday activities and express their dissatisfaction if this connectivity is not available. On the other hand, not only the popularity of such devices increases, but also new applications from different areas such as video, social networks, or cloud services emerge. These aspects together have led to a strong but continuous growth of wireless traffic over the last years that in turn increases the probability of throughput bottlenecks in the near future.

From the perspective of wireless access networks, the key issue is how to resolve or even to avoid such throughput bottlenecks. This is directly related to the aspect of what capacity in terms of users with a certain type of service an access network can offer, and how much of this capacity is still available at a given instance in time. For various wireless technologies, the capacity of a network is not just constant, but depends on how associated devices use the available wireless resources in time, frequency,

and space. For example, devices perceiving lossy wireless links usually employ error recovery mechanisms on the basis of a specific wireless medium access control layer (MAC) protocol. Such error recovery further consumes wireless resources, thus reducing the capacity of the wireless network. In other words, the capacity is influenced by the joint mix of traffic from all the devices that are associated to a given wireless access network. Now, the issue how a device actually selects a network is usually not specified by the technological network standards and is thus dependent on the vendor-specific implementation inside the communication equipment. In addition to that, rather simple approaches regarding such a network selection are frequently applied today, aiming at an association that results in a robust wireless link. For example, associations in WLANs usually just base on received signal strength measures. This however does not include information about how much of wireless resources will be actually occupied by a device, how it will impact other devices present in the cell, and how it affects the overall capacity of the access network. In summary, if associations are not maintained in a *resource-aware manner*—i.e., considering how the wireless channel is used in time, frequency, and/or space—the whole cell as well as other devices may suffer from a degraded network performance regarding throughput, packet delays and losses.

Now, heterogeneity in terms of wireless access technologies does not only enable wireless connectivity for devices anywhere and anytime, but further offers the option to choose among different technological possibilities if a wireless device is within the coverage of multiple wireless networks. In a further step, this conceptually enables to switch an on-going traffic stream of a device from one wireless access to another. This switching is referred to as a *handover*. If the traffic stream is shifted between different wireless technologies, we refer to this as a *vertical* or *heterogeneous handover*. Naturally, the need for such a handover appears if a device with an on-going traffic stream is moving and, as a result, is leaving the coverage of one wireless technology. In such scenarios, mobility-oriented handovers may prevent a loss of connectivity for a mobile end device by switching its traffic stream to an alternative, heterogeneous wireless network that is available.

Besides their context from mobility scenarios, handovers are also used as a tool for other situations. Even when devices remain nomadic, which means that they are not moving during a communication session, a need for a handover may emerge. From the perspective of the end user, this may appear if a network is not able to deliver a desired performance, e.g., in terms of throughput, such that the quality of a running service does not reach a desired level. This in turn may be realized by switching the traffic of the device to an alternative access technology, which is expected to improve the quality of the given service. In contrast, from a network's point of view, one may think about a re-organization of a wireless cell by conducting handovers for selected devices. For example, a network operator may re-evaluate previous association decisions of devices. For a certain subset of devices that performs poorly under the current conditions in the network, the operator may evaluate alternative wireless access options. Following these directions, decisions for handovers are becoming *resource-aware*. Making a step further, dynamic decisions about the access technology to be used (e.g., pushing the users back and forth from a WWAN to a WLAN while maintaining an on-going traffic stream) seem

to be a promising approach to avoid or to resolve throughput bottlenecks in crowded and highly-loaded wireless access networks.

Throughput bottlenecks have been reported specifically in cellular 3GPP WWANs as a result of the traffic increase. Today's research and standardization efforts, specifically from the 3GPP area, aim to *offload* data traffic from WWANs to other wireless accesses such as WLAN or femto cells. Today, at an initial stage, rather simple approaches just make offloading decisions right at the beginning of a communication session. This goes in line with the fact that end users typically prefer WLAN over WWAN access for data traffic, whenever it is available. Such simple offloading approaches just utilize additional capacity in the form of alternative heterogeneous accesses. Thus, offloading or handover strategies aiming for a more sophisticated allocation of end devices may better exploit wireless resources, as heterogeneous networks actually offer additional gains. It is important to understand that heterogeneous networks do not only offer simple capacity improvements due to a variety of networks operating in different frequency bands, but may also result in gains stemming from the fact that end devices may perceive quite different load levels and link conditions in each of the wireless accesses. This can be utilized by jointly reallocating devices to available wireless accesses, i.e., such that the joint capacity of a heterogeneous network and not just the capacity of each access part maximizes.

In summary, handovers as a tool for capacity-based optimizations may further improve the overall situation in the involved wireless accesses. Let us separate the challenges regarding resource-aware handovers in three different dimensions for the following discussion: decision criteria, mechanisms, and policies. *Decision criteria* help identifying when (in time) and where (in parts of a network) certain actions (such as a handover) are required. Further, we refer to a *mechanism* as a specific set of harmonized actions conducted on one or more pieces of network equipment operating jointly according to a given rule. In line with this, we understand a *policy* to be a set of such pre-defined rules, formally describing one or more selected objectives. In other words, a policy represents a strategy, while one or more mechanisms on the basis of the decision criteria actually help realizing a given policy.

Generally, we can think about infinitely many classes of handover policies each following a set of possibly contradicting objectives. Just to name a few examples, one may target to maximize the utilization of resources, to maximize the quality of a given service, or to minimize the energy consumption of network components. In this thesis, we specifically focus on classes of policies that aim to improve the usage and the utilization of wireless resources. We assume that such policies rely on information about how a wireless channel is used in time, frequency, and/or space and denote them as *resource-aware handover policies*.

Challenges for the realization of resource-aware handover policies appear in different contexts as a preparation and an execution of a handover involves different *mechanisms*. They range from a neighbor discovery of alternative networks, the decision making and signaling, the link establishment with a candidate network, a triggering of the handover, to a timely re-establishment of the Internet Protocol (IP) connectivity and the rerouting of user traffic. Mechanisms covering solutions and frameworks for conducting as well as

supporting handovers have been extensively studied and standardized within the Internet Engineering Task Force (IETF), the 3GPP, as well as the IEEE on different layers.

While each family of mechanisms is a research field on its own, we focus in this thesis more specifically on challenges regarding mechanisms supporting a preparation of handovers. This includes the neighbor network discovery as well as the link setup in an alternative access network. We understand both as an important precondition for handovers, as a switching of traffic streams can be only conducted, if the involved wireless devices have alternative access options that are suitable to deliver a certain QoS level. The issue with mechanisms for a preparation of a handover is that researchers usually assume that one network interface card (NIC) is available for each access technology within an end-user device. This simplifies a lot the preparative mechanisms as they can be performed in parallel to the on-going communication via a second NIC, which comes however at an increased cost regarding the consumed energy. By contrast, in reality nowadays, there is a strong trend to limit the number of separate NICs within a single device due to space and cost issues. As a result, multi-standard NICs become a solution of choice which can support different technologies, but may not allow for a simultaneous use of multiple wireless links in parallel. We refer to such a NIC as a *multi-mode radio* if at a given instance of time only a *single* wireless access link to one of the technologies is possible at all. How to enable a usage of heterogeneous links over such multi-standard NICs while meeting QoS constraints of an on-going transmission, including constraints imposed by real-time voice connections, is a key challenge in the area of a preparation for handovers.

Further challenges towards a practical realization of resource-aware policies appear in terms of *decision criteria and related technological parameters*. There, one challenge lies in a construction of realizable decision criteria. More and more, related work on handover decision mechanisms tends to be based not just on a single parameter, e.g., decisions based on the received signal strength of a certain link. Approaches from the literature already analyzed the impact of various different parameter sets thereby also focussing on mathematical approaches to combine a multitude of parameters. The challenging part for our resource-aware policies is to reflect the underlying technological behavior that impacts how consumed wireless resources are exploited by each individual device. While mutual information about this behavior can be included in a large set of technological parameters, the actual challenge lies in the *formation of a unified performance metric* allowing a direct comparison of devices on a single scale in given technology on the basis of *available* and *suitable* parameters.

Another challenge lies in the *availability of parameters*. By availability we mean that measures of technological parameters are accessible on a network component that computes the performance metric. Basically, we can divide these parameters into three categories according to an accessibility at different network components. For a brief discussion, let us assume that we aim to gather the parameters on a selected piece of network equipment. A first subset of technological parameters may be directly measurable and accessible there, e.g., on a WLAN access point (AP) of an access network. In contrast, a second subset of parameters may instead not be directly available because it is measurable only at a disjoint piece of network equipment, e.g., on an end-user de-

vice. Lastly, the third subset consists of 'hidden' parameters which are only used inside internal modules of a network equipment. Typical representatives of the latter category are parameters that belong to closed, vendor-specific algorithms of a network equipment and are not exposed via typical management interfaces. The second category, where parameters are available only on disjoint network components, effectively calls for an appropriate orchestration of parameter measurements as well as their signaling between involved network components. Standard amendments such as IEEE 802.11k/v from the WLAN arena tackle these aspects of the second category, thus this issue is not that critical although still an appropriate selection among the standardized measurement parameters has to be made. Contrary, a vast group of challenges appears in the context of the third category, where technological parameters are not directly accessible in either of the network components. If these parameters are critical with respect to a realization of a certain performance metric and its related decision criteria, they also become critical regarding an implementation of the specific policy.

Further, a set of parameters can only become available *after* related measurements have been completed. This allows to obtain parameters and calculate performance metrics in a reactive fashion only. Then however, it is not possible to *predict* how a traffic stream, e.g., being subject of a handover to WLAN, will actually behave in its target wireless access cell over a certain time period. Such a prediction of the behavior may consider parameters regarding the mobility of a terminal, its traffic pattern, and its resource consumption. Considering WLANs, devices with a strongly fluctuating channel and increased collision levels degrade the whole performance of a cell, such that handover decisions based on proper estimates of parameters could dramatically avoid impairments for all clients in a proactive fashion.

Finally a further set of challenges for policies, mechanisms, and decision criteria is related to the cooperation level of network owners. The critical issue with today's wireless networks is, both from the 3GPP as well as IEEE standardization bodies, that they do not yet support such resource-aware handovers in different directions pari passu. Usually, policies are defined and decisions are conducted in a centralized manner by WWAN operators, such that handovers from WWANs to WLANs do not consider the viewpoint of WLAN operators at all. Further, handovers from WLANs back to WWANs are not yet fully supported by WWAN operators. However, such equitable decisions may be of great importance especially for those WLAN operators who are required to cover and support a large group of end users, e.g., in enterprise networks or in hotspot scenarios appearing in cafes, malls, or large environments such as halls inside train stations or airports. As a result, the position of WLAN operators or owners remains quite weak in the context of the current offloading discussions, where it is assumed that all data traffic of an end device can be simply just moved from WWANs to WLAN cells.

1.1 Thesis Contribution

Wireless access networks do not yet support resource-aware handovers in all the different dimensions regarding related policies, mechanisms, and decision criteria. More recently,

resource-aware handovers have emerged in the context of the offloading discussion, where data traffic is shifted from WWANs to WLANs. While there remain challenges in each distinct wireless technology, we focus on these aspects specifically from the WLAN perspective, as the popularity and the availability of these networks make them a superior candidate for vertical handover scenarios. As a basis for our work, we consider in addition to WLAN hotspots to have a heterogeneous WWAN using a complementary technology in terms of coverage, mobility support and QoS for the end user. Without loss of generality, the WWAN may consist of the third and fourth generation of mobile telecommunications technology (3G, 4G) or even of mobile WiMAX, however covering completely the WLAN hotspots. In addition, to fully exploit the gains of heterogeneous WLAN and WWAN accesses, we further aim at a support of handovers not only from WWAN to WLAN networks, but also in the reverse direction. In the analogy to offloading, we refer to a handover from a WLAN back to a WWAN as *onloading*. In the following, we briefly survey the contributions of this thesis in the context of resource-aware handover policies, related mechanisms and criteria for such an onloading.

Opportunistic Preparation of Handovers For an onloading handover, the support of a timely preparation and the actual switching of a traffic stream is a crucial issue. Thereby, the preparation phase includes a collection of different mechanisms ranging from neighbor network discovery and link setup to the IP path establishment via an alternative access network. On-going communication and the handover preparation in an alternative network in parallel imposes a challenge if only a single NIC is available in an end-user device. Thus, the first contribution of this thesis addresses the issue *how to enable a usage of multiple links over a multi-mode NIC for a support of handover preparation mechanisms*. For this, we design an innovative approach from the WLAN perspective that we denote as *opportunistic preparation of handovers*. There, an end device only pauses its on-going WLAN communication for extremely short but frequent time spans. During each of these time spans, the end device conducts handover preparation steps such as scanning for other technologies. We select the duration of these time spans to be small enough such that they do not violate the QoS constraints of the application being used. Further, to avoid packet losses or even a break of the wireless connectivity, we apply selected IEEE 802.11 power save mechanisms, effectively allowing us to pause WLAN communication during the short time spans. We analyzed and evaluated our general approach regarding two selected application examples covering homogeneous WLANs as well as heterogeneous WLAN/WiMAX networks. Thereby, our results identify the limits regarding timing dependencies and background load as well as the range of the signaling overhead induced by our approach.

Handover Candidate Selection in WLANs Today, an end-user device usually tries to associate and authenticate with a WLAN network, either when it detects that it is within the coverage of a preferred hotspot, or the end user manually forces the device to switch to WLAN. On the one hand, end users strongly favor WLAN over WWAN access for data traffic, because WLAN connectivity comes mostly for free and offers higher

throughput peak rates. On the other hand, such simple technology selection scheme completely neglects the question whether the device with a certain traffic stream is actually suitable for a certain WLAN hotspot, i.e., whether it is not negatively affecting the cell. On MAC and PHY level, a whole 'zoo' of WLAN parameters is available, e.g., measures of the received signal strength, the required number of retransmissions, the selected MCSs for transmissions, or contention level parameters, reflecting the operational point of the WLAN hotspot. From these parameters, one can easily obtain the time that the wireless channel is occupied by a given traffic stream, commonly denoted as the 'airtime'. Nevertheless, it not easy to derive from this which traffic stream out of many is actually negatively influencing the whole cell. However, for a maximum loading of the cell, a WLAN operator is required to identify unsuitable traffic streams that should be selected for a heterogeneous onloading handover back to the WWAN. The term 'unsuitability' thereby refers to traffic streams with negative impact on the capacity of the wireless cell. We refer to this as the *selection process of potential handover candidates*. For a clear distinction between suitable and unsuitable traffic streams, this thesis firstly presents a performance metric aiming at an efficiency evaluation in terms of the occupied resources, i.e., the airtime. Our performance metric easily aggregates selected WLAN MAC and PHY parameters into a unique measure, such that it covers detailed technological insights with only low computational effort. We evaluated our metric in an exemplary scheme for the selection of handover candidates in a WLAN cell to be onloaded to the WWAN by comparing the gains of our approach with the classical RSSI-based as well as simple random decisions.

WLAN Link Data Rate Estimation One of the problems for resource-aware handovers lies in the fact that one is usually not able to predict how a certain traffic stream will behave in the target network. For a wireless device initiating the association process with a certain cell, its future airtime and its impact on other wireless devices is completely unknown. As a result, it is impossible to judge the future behavior of this device. Considering WLAN systems, manifold reasons for the problem of unknown airtime exist. For an arriving device in a WLAN cell that has recently completed its association procedure usually the following issues are unknown: its traffic pattern, the effect of the traffic pattern on the collision level in the WLAN cell, its channel state, as well as its scheme for the selection of link data rates (i.e., the applied modulation and coding scheme, MCS). To make it even worse, the MCS selection for WLAN transmissions depends on the proprietary algorithm of each WLAN card vendor, potentially leading to a diverging behavior of users with WLAN cards from different manufacturers. As specifically transmissions with low-rate MCSs may degrade the capacity of a WLAN cell, we focus on an estimation regarding the data rate selection behavior of a device in a third contribution of this thesis. Thereby, it is important to mention that we do *not* predict the traffic pattern of a device. Instead, we rely on devices with stationary traffic and assume to have a priori knowledge about these traffic patterns.

Accordingly, this thesis presents our innovative *data rate estimation scheme 'DARA'* for WLAN networks. The core mechanism of DARA estimates the rate selection of a

WLAN end-user device just by observing its behavior on short time scales—without having any knowledge about the applied rate adaptation algorithm. For the estimation, we present the selected machine learning approach used by DARA and demonstrate its efficiency using both simulated WLAN configurations as well as measurements from a WLAN setup in an on-campus office environment. We studied DARA's estimation accuracy for the selected rate adaptation (RA) schemes 'adaptive auto rate fallback (AARF)' / 'adaptive multi rate retry (AMRR)' and 'Minstrel' with data from simulations as well as measurements. Finally, we utilized DARA's estimates as a basis for selecting suitable devices running voice over IP (VoIP) traffic in WLAN hotspots and to onload unsuitable devices back to a WWAN instead.

1.2 Structural Overview

The thesis is structured as follows:

Chapter 2, Infrastructure Wireless LAN Basics: First, we survey the IEEE 802.11 standardization regarding the typical infrastructure architecture, as well as management issues such as network synchronization and the network entry. Then, we explain the basic PHY layer variants, present the functionality of the MAC protocols, describe their issues for QoS constrained traffic, and present related MAC extensions. Finally, we introduce the basic 802.11 power management mechanism.

Chapter 3, Resource-Aware Handovers: This chapter gives an overview about the offloading context, provides a taxonomy of heterogeneous handovers and includes a survey about architectures and frameworks supporting handovers from 3GPP and IEEE standardization bodies.

Chapter 4, Challenges and Scope of Thesis: Firstly, we discuss the challenges regarding resource-aware handovers in detail. Further, we present our reference scenario together with our basic assumptions and present the underlying architectural framework based on various extensions from IEEE standardization bodies. From this, we derive the primary scope of the thesis by selecting the most important directions for an optimization from the WLAN perspective.

Chapter 5, Opportunistic Preparation of Single-Radio Handovers: For our device-centric extension regarding a support of handover preparations, we first present our general design highlighting the means of the IEEE 802.11 standard being utilized for a practical realization of our approach. Then, we present two specific application cases of the general design, discuss their features, and present results of their evaluation.

Chapter 6, Handover Candidate Selection: We first argue how an 'onloading' of WLAN traffic back to WWAN may improve the overall situation of both WLAN owners

as well as WWAN operators. Then we present the design of our selection metric by discussing the goals for each of its components in detail. An initial test case intuitively shows the behavior of the metric by considering a simplified scenario. The comprehensive performance evaluation by means of simulations finally compares our scheme with others highlighting the advantages of our 'onloading' metric.

Chapter 7, Link Data Rate Estimation: This chapter first gives an overview about the basic functionality of the considered estimation scheme and describes the selected model from the machine learning area. We present estimation results both from simulations as well as measurements in our WLAN testbed. Finally, we use the data rate estimates in a simulative performance evaluation identifying the gains if applied for admission control decisions in a WLAN hotspot.

Chapter 8, Conclusions and Outlook: Finally, we summarize our work, emphasize the conclusions of each contribution and end up with an outlook towards future directions regarding traffic off- and onloading.

Chapter 2

Infrastructure WLAN Basics

Since its first version back in 1997, the IEEE 802.11 standard[1] specifies *medium access control (MAC)* and *physical layer (PHY)* mechanisms for WLANs. First, this chapter starts with a basic discussion about the WLAN architecture for infrastructure networks. Then, we describe the synchronization process inside a WLAN cell, before we introduce how a device becomes a member of a WLAN network. Afterwards, we survey existing PHY variants and their frame structures, present related backward compatibility issues and their consequences, and argue about the need for a link data rate adaptation. In a next step, we focus on MAC level tasks that appear throughout this thesis. For a basis of the legacy 802.11 medium access, we first present the underlying mechanisms, before we finally explain the *distributed coordination function (DCF)* and the *point coordination function (PCF)*. We end with a discussion about quality of service (QoS) related issues regarding the basic WLAN medium access and describe the standardized solutions of the 802.11e amendment.

In certain parts of this chapter, we follow the description of the well-known book "Mobile Communications" [1], as it includes from our perspective the most intuitive presentation of the WLAN basics. The interested reader is referred to the complete chapter about WLANs [1, Ch. 7, pp. 161–214]. We really focus on the basics of WLANs in the following sections. For all the sophisticated concepts of more recent standardization activities inside 802.11, we refer the experienced reader to the survey in [2], while the "IEEE 802.11 Handbook" [3] nicely highlights all the details about older 802.11 variants until 2005. We note that some parts of this chapter have published before in [4, 5].

2.1 WLAN Architecture

The 802.11 standard [6] introduces the notion of a *WLAN station (STA)*. A STA can be either simply an end device or an *access point (AP)*, whereby the latter provides means

[1] For the sake of convenience and readability, we skip the term 'IEEE' from now on whenever we refer to the 802.11 standard, an 802.11 amendment, or an 802.11 working group

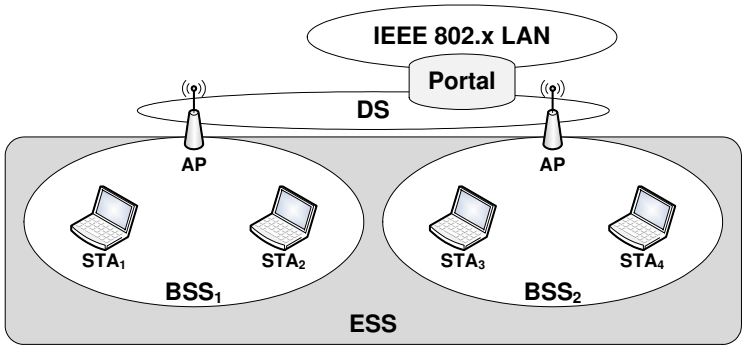


Figure 2.1: Basic WLAN infrastructure network, modified after [6, Fig. 4-3, p. 48] and [1, Fig. 7.3, p. 169]

for the connectivity to other (wired) networks.[2] In *infrastructure WLANs*, STAs are connected to such APs. Thereby, a single AP, together with its connected STAs, forms a *basic service set (BSS)*. Further, multiple BSSs may be interconnected, usually by a wired network, thus forming an *extended service set (ESS)*. Such an ESS is setup only within one administrative domain under the control of one WLAN network operator. ESSs are quite common in large company- or campus-wide WLAN networks being denoted as *enterprise networks*. The part interconnecting multiple BSSs is referred to as distribution system (DS), while the entity connecting the DS to other networks is known as the *portal*. All components are visualized in Fig. 2.1.

2.2 WLAN Protocol Stack

STAs and APs include the complete 802.11 functionality of PHY and MAC level following the International Organization for Standardization (ISO) / Open Systems Interconnection (OSI) layered concept [7]. In addition, every 802.11 device has a management plane that includes a *physical layer management entity (PLME)*, a *MAC sublayer management entity (MLME)*, and a *station management entity (SME)* as shown in Fig. 2.2. Basically, the PLME is responsible for providing access to PHY-related management functions detailed later and to enable the configuration of the PHY. The MLME covers management tasks like synchronization, network entry, and power management.

Configuration, management, and status information for PHY and MAC are stored in *management information bases (MIBs)*. Thereby, a MIB is a structure of different, predefined pieces of data representing parameters and variables such as counters. Both, PLME and MLME have their own MIB. Finally, the SME acts as an interface between higher layers and both, PLME and MLME. By this, the SME provides a service to other

[2] In contrast to the 802.11 standard, we follow the notion of a STA (as a non-AP STA) throughout this thesis. Further, we apply the terms *STA*, *node*, and *end device* interchangeably. Else we use explicitly the term AP to avoid any confusion.

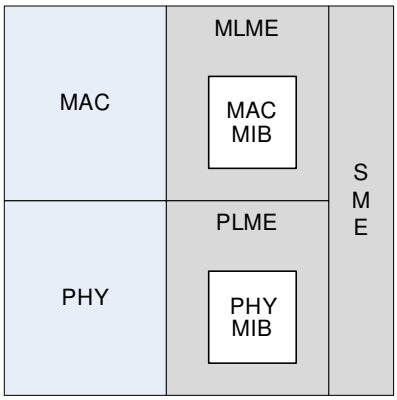


Figure 2.2: WLAN protocol stack, modified
after [6, Fig. 6-1, p. 104]

layers allowing to access the PLME and MLME MIBs. This is realized by *get* and *set* commands enabling to read-out or to change pieces of data in the MIB structure.

2.3 The Basic Time Synchronization Concept

802.11 follows a simple concept for synchronizing the AP with its STAs. As will be discussed further below in more detail, to keep all members of a BSS synchronized is a fundamental basis for various issues such as the wireless medium access as well as the support of power save mechanisms.

In infrastructure BSSs, the basic synchronization of all the connected STAs is realized by a centralized clock mechanism. Loosely speaking, the AP serves as a centralized master clock and distributes its notion of time to its connected STAs. For this, each STA and each AP possesses a local *timing synchronization function (TSF)*, which is a counter that increases on a per microsecond basis. Now, for a synchronization, the AP regularly schedules to send out an update of its current time announced in so-called *beacon* frames in a broadcast fashion. Each STA that has received a beacon adjusts its local TSF to the received time value of the AP. The point in time at which the upcoming beacon shall be transmitted by the AP is denoted as *target beacon transmission time (TBTT)*. The time span between two consecutive TBTTs is known as the *beacon interval* being usually around 100 ms in practical systems.

It may happen, however, that the AP is not able to transmit its scheduled beacons exactly at each TBTT, because another STA is yet conducting a transmission on the channel. According to the medium access rules detailed later, the AP then has to wait until the wireless medium becomes free again, effectively leading to *late beacons*. To circumvent that a late beacon affects the synchronization inside a BSS, the AP always

obtains a current timestamp from its TSF immediately before the actual transmission and includes this value in the beacon frame.

2.4 WLAN Network Entry

To become a member of a BSS and to obtain WLAN connectivity, a STA initially has to conduct a set of distinct steps that are denoted as *WLAN network entry process*. It consists of the network discovery, the authentication, and the association phases. For a brief description of each of these phases, let us consider a WLAN STA that has been just powered on in a given environment being covered by a WLAN AP.

Under such circumstances, our STA first of all has to search for available WLAN APs. This mechanism is known as *network discovery*. In WLANs, this search is also denoted as *scanning* as the STA sweeps through the WLAN channels checking for available BSSs. Two different flavors are available: *active and passive scanning*. In the first, a WLAN STA explicitly sends out probe request frames on each channel, to which APs answer with probe response frames. In contrast, the passive scanning only listens to each channel passively, identifying present APs just by receiving their beacons. The scanning does not only give the presence of APs on channels, but further conveys information about each BSS, as included in the beacon as well as the probe response frames. This includes details about the PHY-level flavor, the parameterization of the BSS, as well as specific requirements, e.g., whether the AP offers open access or presupposes an authentication of STAs. Note that the 802.11 standard does not specify whether a STA scans all available WLAN channels or just a subset. In addition, the sequence of the channels to be scanned is left open for vendor-specific choices.

From the scanning process, the STA obtains a list of candidate WLAN APs from which it selects one. How this selection is actually done, is again out of scope of the 802.11 standard. Nevertheless, the most common selection schemes rely on the perceived signal strength of the beacon or probe response frames arrived at our STA [8]. In a next step, our WLAN STA has to undergo the *authentication* which is intended to ensure that the STA entity is allowed to access the BSS. The exact sequence of frame exchanges between STA and AP depends on the specific authentication flavor used inside the BSS. For the details about these authentication mechanisms, the interested reader is referred to the surveys in [9,10]. What is important for our basic discussion here is that the STA initiates the authentication with a request frame and after completion, the AP indicates the success or failure of the authentication in a response frame.

In the third phase, the STA has to accomplish the *association* phase starting again with a request frame. The AP answers with a response frame, finally indicating whether its grants or denies access to the BSS for the STA. Further, this handshake is used to set various parameters being important for the behavior of the STA in the BSS, such as the set of link data rates, contention parameters (for the enhanced distributed channel access, EDCA, explained below), and others.

After these three phases, the WLAN STA is a member of the BSS and is allowed to conduct transmissions of all types of frames. For the sake of completeness, we note

that some of the authentication procedures from the second step have been shown to be vulnerable. Therefore further means have been incorporated with the 802.11i amendment in 2004 [11], not only improving the authentication in the second step, but also offering the possibility for a second level of authentication between the STA and the AP (or some other entity in the DS) on top of the basic WLAN network entry. With this option, although a STA is associated, all its user traffic is blocked by the AP until the STA has successfully completed this extended authentication procedure. Again we refer for the details to other surveys [9, 10].

2.5 Physical Layer Basics

Roughly speaking, the PHY layer is responsible for the transmission of digital data that arrives in the form of frames from upper layers. For this, the PHY basically converts the digital data into analog signals to be transmitted on the physical medium. The whole process for mapping digital data to a representation of analog signals in the form of *symbols* is known as modulation and coding. For this, a modulation and coding scheme (MCS) maps a fixed number of N bits to a symbol. We note that on the one hand, the more bits an MCS maps to a symbol, the higher becomes the actual *link data rate* of a transmission as the amount of encoded bits in a symbol with a fixed duration increases. On the other hand, a higher number of bits per symbol requires a higher signal-to-noise ratio (SNR) at a receiver for an error-free reception of a transmission. Finally, for an emission of a wireless transmission in the form of electromagnetic waves, a transmitting PHY conducts a carrier modulation of the symbols thereby shifting them by a selected scheme in the frequency dimension from the base band to the actual frequency band in use by given technology.

In the context of WLANs starting back in 1997, when the original 802.11 standard [12] appeared, to its most recent version in 2012 [6], a diversity of different PHY specifications was included. We give a brief overview about these different variants mainly following the survey of Hiertz et al. [2]. Thereby, we highlight the frequency bands in use, the applied carrier-modulation techniques, as well as the set of available MCSs resulting into various raw link data rates. Afterwards, we discuss backward compatibility issues of the PHYs and related consequences, before we review practical considerations regarding a selection of MCSs for transmissions.

2.5.1 Overview about Physical Layer Variants

Already the first standard version from 1997 offered different PHY specifications, one for infrared communications in the 316 to 353 terahertz (THz) and two for the 2.4 gigahertz (GHz) industrial, scientific, and medical (ISM) band, whereby the infrared PHY never made it into commercially available products [2]. For the 2.4 GHz band, the 1997 standard introduced two different specifications, the first using *frequency hopping spread spectrum (FHSS)*, while the second relied on *direct sequence spread spectrum (DSSS)*. Both techniques support MCSs resulting into raw link data rates of 1 and 2 megabits per

second (Mbps). Aiming to make wireless transmissions more robust against interference, the spread spectrum approaches enable a support of neighbor WLAN networks which have overlapping coverage areas. While FHSS applies a randomized hopping sequence for each network on a high number of small 1 MHz channels, DSSS instead relies on larger 22 MHz channels and spreads the wireless signal in the frequency domain with a given code sequence [1]. In 1999, 802.11b [13] amended the DSSS specification by introducing a *complementary code keying (CCK)* based modulation technique, leading to an additional support of 5.5 and 11 Mbps link data rates.

In contrast, 802.11a [14] introduced the concept of *orthogonal frequency division multiplexing (OFDM)* in the 5 GHz band for 20 MHz channels. OFDM enables the parallel transmission on 48 orthogonal data subcarriers, each with a bandwidth of 312.5 kHz. Consequently, 802.11a is capable to support raw link data rates of 6, 9, 12, 18, 24, 36, 48, and 54 Mbps. As a result of its specification for the 5 GHz band only, the 802.11a operation did not allow a backward compatibility with older WLAN devices. In 2003, 802.11g [15] finally introduced the OFDM concept of 802.11a in the 2.4 GHz band, thus being backward compatible with the older 802.11/b amendments, while at the same time offering raw link data rates up to 54 Mbps.

In 2009, the 802.11n amendment introduced "Enhancements for Higher Throughput" [16], both for 2.4 and 5.0 GHz. These enhancements include *multiple input multiple output (MIMO)* with up to four antennae on sender and receiver side as well as an optional channel size of 40 MHz. The MIMO feature can be used for *spatial diversity* as well as for *spatial multiplexing* [17]. While the first aims for a support of robust transmissions, the second targets an increase in throughput. We refer the reader to the editorial survey note in [17] for a broader discussion of the MIMO features. Overall, 802.11n specifies 77 MCSs in total, leading to raw link data rates of up to 600 Mbps [2].

2.5.2 Frame Structures on PHY level

We briefly review the framing concept on PHY level and the resulting frame structure for different 802.11/b/g PHY modes [6]. Generally, a PHY frame consists of a *physical layer convergence procedure (PLCP) preamble* and *header* plus the *PLCP service data unit (PSDU)* which contains the headers and the data of upper layers as shown in Fig. 2.3. The PLCP preamble synchronizes the hardware of the receiver and signals the start of a PHY frame. Among others, the PLCP header includes important information being required for the decoding of the frame on the receiver side such as the applied MCS and the length of the total frame in time. PLCP preamble and header are both transmitted at the most robust MCS resulting in the lowest possible link data rate for the corresponding PHY configuration. In contrast, the PSDU part may be transmitted at any of the MCSs supported by a PHY.

Fig. 2.3a) shows the mandatory long preamble for 802.11 DSSS and 802.11b *high-rate DSSS (HR/DSSS)* being transmitted with *differential binary phase shift keying (DBPSK)* which results in 1 Mbps. This leads to a duration of 192 μs for PLCP preamble and header. Further, an optional frame format of 802.11b HR/DSSS reduces the PLCP

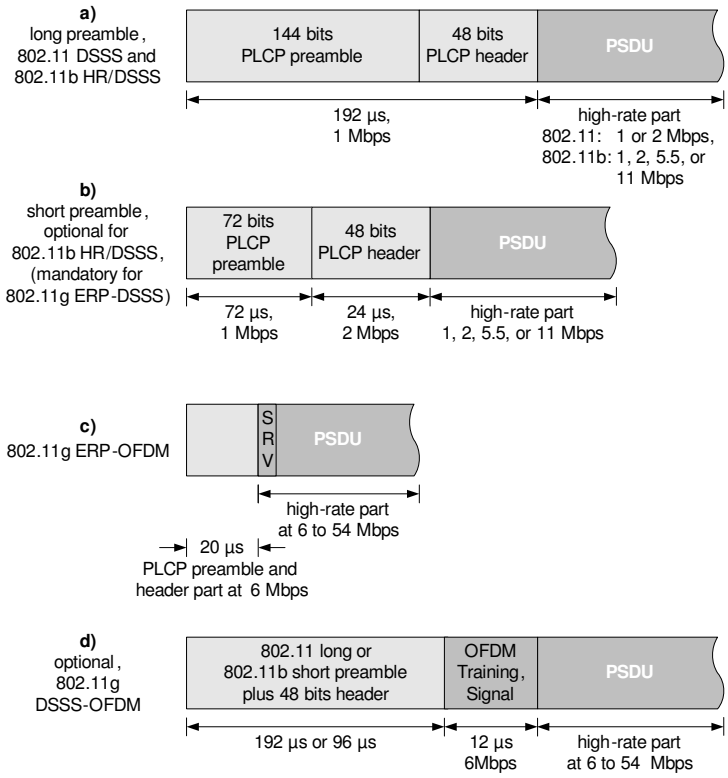


Figure 2.3: Framing of 802.11/b/g, a) to c) showing individual PHY modes, d) giving the backward-compatible DSSS-OFDM mode of 802.11g, modified after the basic framing figures given for each PHY in [6]

preamble and header in time down to 96 μs (Fig. 2.3b). While 802.11 DSSS supports 1 and 2 Mbps for the transmission of a PSDU, 802.11b HR/DSSS allows up to 11 Mbps.

Compared to mature 802.11/b setups, 'pure' 802.11g reduces the size of the PCLP preamble and the part of the PLCP header transmitted at the most robust rate in its *extended rate PHY OFDM (ERP-OFDM)* mode down to 20 μs. This results in a PHY frame shown in Fig. 2.3c). With *quadrature phase shift keying (QPSK)* resulting in 6 Mbps, this PHY mode use a higher-rate MCS as the basic, most robust rate, but is not interoperable with 802.11 DSSS and 802.11b HR/DSSS. The high-rate part starts for 802.11g ERP-OFDM with a 16-bit service field (SRV), which initializes the descrambler on receiver side and by standard definition belongs to the PLCP header, followed by the data symbols including the PSDU that are transmitted at rates from 6 to 54 Mbps.

Leaving older, obsolete PHY modes aside [6], 802.11g does not only offer ERP-OFDM, but also specifies the optional DSSS-OFDM and the ERP-DSSS mode for interoperability issues with 802.11/b. As shown in Fig. 2.3d), the DSSS-OFDM PHY applies the older 802.11b framing with long or short preambles and transmits the PSDU

parts with 802.11g link data rates. The latter requires some short training symbols for OFDM, a signal field giving the selected MCS and the duration of the PSDU, and a 6 μs signal extension at the end (not shown in the figure). Since DSSS-OFDM is an optional PHY mode, it may however not be supported by all 802.11g NICs. Further, the most recent 802.11-2012 standard marks DSSS-OFDM as depreciated. Lastly, the 802.11g ERP-DSSS PHY mode reuses the 802.11/b framing and their MCSs, whereby it has a mandatory support of the short preambles. In other words, ERP-DSSS results in the same frame structure as 802.11 DSSS and 802.11b HR/DSSS shown in Fig. 2.3a) and b).

For a comparison, Table 2.1 surveys the slot times, the PLCP preamble plus header duration and raw data rates resulting out of the different MCSs of each PHY amendment.

2.5.3 Backward Compatibility

During the standardization of new PHYs for the 2.4 GHz bands, the 802.11 working group always aimed to be backward compatible with earlier PHY specifications. On the one hand this ensures that mature WLAN NICs are still able to operate in hotspots offering more recent PHY capabilities, such that consumers are not forced to buy new equipment. This is one of the reasons that WLAN technology has been able to reach its vast popularity. On the other hand, it is this backward compatibility that imposes negative implications on the capacity of a WLAN cell as it increases the duration both for the medium access as well as a transmission. This has several reasons. First, the WLAN PHY specification determines a smallest time granularity for the medium access. This time granularity is also known as a slot time. For newer PHYs not being in backward compatible mode such as 802.11g/n it is 9 μs small, less than half of the size for older 802.11/b. As these slots are applied for the timing of the medium access regarding every data frame, certainly their size has an effect on the overhead of a transmission.

Second and more importantly, PHYs specify various options to allow an interworking with mature 802.11 devices, which however affect the transmission duration of a data frame. Backward compatibility specifically enables that also mature STAs in a BSS can at least detect and identify on-going transmissions of WLAN equipment supporting more recent standard amendments. In the following, we briefly review this aspect both for 802.11b to 802.11 as well as for 802.11g to 802.11/b compatibility. As 802.11b HR/DSSS directly supports the long preamble format, it is already backward compatible to legacy 802.11 DSSS. Nevertheless 802.11b must not use its optional, short preambles together with old 802.11 devices. Thus, whenever one or more mature STAs are associated with a certain BSS, the whole operation has to be adapted accordingly.

To handle inter-operability with mature 802.11/b devices, 802.11g basically defines three different options, which base on the PHY modes and their framing that we discussed in the previous section. The first option for backward compatibility just uses the DSSS-OFDM mode that applies the older 802.11b framing with long or short preambles and transmits the PSDU parts with 802.11g link data rates (c.f. Fig. 2.3d). In contrast, the second and third option base on the ERP-DSSS PHY mode that applies the 802.11/b framing. The second option further reuses a MAC mechanism, which is denoted as request-to-send (RTS)/clear-to-send (CTS) handshake. While we describe this mech-

Table 2.1: Survey of 802.11 PHYs and their link data rate sets

Amendment/PHY	slot time [μs]	PLCP preamble and header [μs]	link data rates [Mbps]
802.11 DSSS	20	192	1, 2
802.11b HR/DSSS	20	96/192	1, 2, 5.5, 11
802.11g ERP-OFDM	9	20	6, 9, 12, 18, 24, 36, 48, 54
802.11g DSSS-OFDM	20	96/192	6, 9, 12, 18, 24, 36, 48, 54

anism together with the MAC basics in detail later in Sec. 2.6.1, we introduce its concept in the context of backward compatibility here. Basically, an 802.11g sender transmits a short RTS frame to an 802.11g receiver, which responds with a CTS frame. Both, RTS and CTS include the duration of a following transmission sequence encoded in high-rate 802.11g ERP-OFDM frames. As the RTS/CTS handshake is transmitted in the ERP-DSSS PHY mode, all 802.11/b/g STAs that can either decode the RTS or the CTS frame know how long the channel will be occupied. Then, just in this time span, the two 802.11g STAs are allowed to use ERP-OFDM for their transmission sequence. Although 802.11/b devices are not able to decode the 802.11g ERP-OFDM signal, they know when in time the transmission sequence will be completed. Finally, to reduce the overhead of a complete RTS/CTS handshake, 802.11g further introduces the third option, which is denoted as CTS-to-self. There, a sending STA just reserves the wireless channel by a single CTS frame addressed to itself, before issuing an ERP-OFDM transmission in the reserved time span. For a comparison of the three options and their influence on the capacity of a WLAN BSS, we refer the interested reader to the work of Vassis et al. [18].

2.5.4 Demand for a MCS Selection

In the 802.11 amendments, a clear tendency appeared towards an increasing number of MCSs. While 802.11b has four MCS options, 802.11g ERP-OFDM comes up with eight MCSs, while 802.11n specifies in total 77 MCSs. From such a set of MCSs, an 802.11 PHY uses the most robust MCS for the transmission of the PLCP preamble and header and any of its MCSs for the PSDU part. Thereby, MCSs for the PSDU part may be varied by a PHY on a frame to frame basis.

As a result, the need for a proper selection emerges. Such a selection of an MCS for the high-rate part of a frame is a decision making problem that suffers from a tradeoff between two aspects. On the one hand, high-rate MCSs result in lower transmission times and higher throughputs. On the other hand, high-rate MCSs require stronger SNRs for an error-free reception. In contrast, low-rate MCSs increase the transmission time but are more robust against transmission errors.

As a matter of fact, the 802.11 standard only specifies the set of MCSs for each PHY, but does not regulate when to use which MCS for a data transmission. As this

aspect is not standardized, each WLAN implementation has to apply an own, proprietary solution, which is known as a *link data rate adaption scheme*.

2.6 Basic WLAN Medium Access

The WLAN medium access is defined by the mandatory distributed coordination function and the optional point coordination function, whereby the latter is actually rarely used because of its complexity, its overhead in terms of signaling, and other drawbacks highlighted in the next section. Now, before we discuss both medium access functions, we first present important prerequisites that are utilized by both.

2.6.1 Underlying Mechanisms

The 802.11 standard makes use of certain underlying mechanisms and concepts. First, we discuss the purpose of immediate acknowledgments and retransmissions, then describe the task of carrier sensing, and finally introduce different inter-frame spaces being used for a handling medium access priorities.

Immediate Acknowledgments and Retransmissions WLAN STAs are capable to either transmit or to receive at a given instance of time, but are not able to support both in parallel. This is denoted as *half-duplex* communication mode. As a result, a sending STA is not able to determine what actually happens during its transmission. In addition, such a capability would not help much, since the wireless signal *as perceived on the receiver side* essentially governs whether a transmission can been successfully decoded there. Many different impairments may happen in WLAN channels, ranging from other WLAN STAs transmitting at the same time, physical phenomena of wireless radio signal propagation as detailed in Appendix B.3, or even interference with other non-WLAN radio technologies operating in overlapping frequency bands. In consequence, the 802.11 standard introduces small, explicit frames that a receiver immediately sends out after a successful reception of a data frame. These frames are denoted as *acknowledgments (ACKs)* and give the sender of a data frame a positive indication of a successful reception.

Note that in case of a missing ACK, a sender of a data frame is not able to obtain any knowledge about the actual source of error(s). In other words, either the transmission of the data frame or of the ACK could have been disturbed by any of the reasons described above. In case of a missing ACK, the sender of the data frame assumes that an error appeared during its transmission. For the error recovery, the sender initiates a *retransmission* of the same data frame and marks it by setting a *retry bit* in the MAC header.

Carrier Sense To avoid that a STA initiates a transmission while the wireless channel is already occupied, 802.11 basically relies on a specific flavor of *carrier sense multiple access (CSMA)*. There, a STA 'listens' (i.e. senses) the channel prior to a transmission as further detailed in the following subsections. For this, WLAN STAs combine two concepts denoted as *physical* and *virtual carrier sense*.

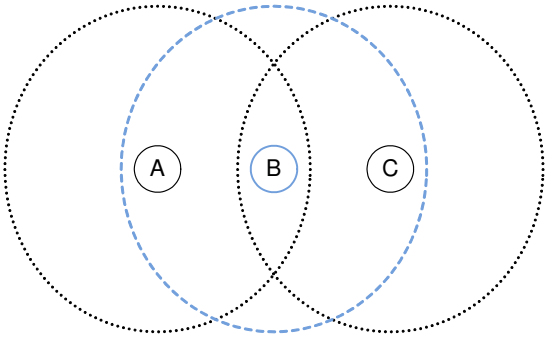


Figure 2.4: Hidden terminal scenario, modified after [1, Fig. 3.1, p. 62]

The basic idea behind the first is that a STA simply senses the channel before it actually starts a transmission. Each WLAN PHY has a feature that detects the energy of incoming signals at its antenna, thus being capable to determine whether the wireless channel is occupied or free. To inform the MAC layer about state changes of the channel, the PHY offers the *clear channel assessment (CCA) function*, which indicates whether the wireless medium has become idle or busy.

The second, the virtual carrier sense, was designed to handle situations where other surrounding STAs are only capable to receive a part of a certain transmission sequence. For example, a third STA may be within the transmission range of a sender but not of its receiver. Thus the third STA is only capable to receive a data frame emitted by the sender but cannot detect the following immediate ACK. Applying only the rules of the physical carrier sense, the third STA would attempt to access the wireless channel after the end of the data frame, thus disturbing the reception of an ACK. To circumvent these situations, the MAC header of each frame includes a *duration field* that specifies the remaining duration of the complete transmission sequence. Each STA being able to decode a frame with a duration field updates a timer, denoted as *network allocation vector (NAV)*, indicating the remaining duration until the wireless channel will be idle again. Thus, all surrounding STAs, even seeing only a part of a transmission sequence, know the point in time until the channel will be occupied by others. Therefore, the 802.11 standard considers a free wireless channel only, if *both* physical and virtual carrier sense indicate an idle channel.

Inter-frame Spaces Further, as a basis for the two medium access functions detailed in the next subsection, the distributed coordination function (DCF) and the point coordination function (PCF), the 802.11 standard introduced timing priorities for the medium access by means of inter-frame spaces which are constant gaps in time between subsequent frames 'on air'. The three basic parameters are the *distributed inter-frame space (DIFS)*, the *PCF inter-frame space (PIFS)*, and the *short inter-frame space (SIFS)*. *DIFS* applies if a WLAN device has a data frame ready to transmit and is

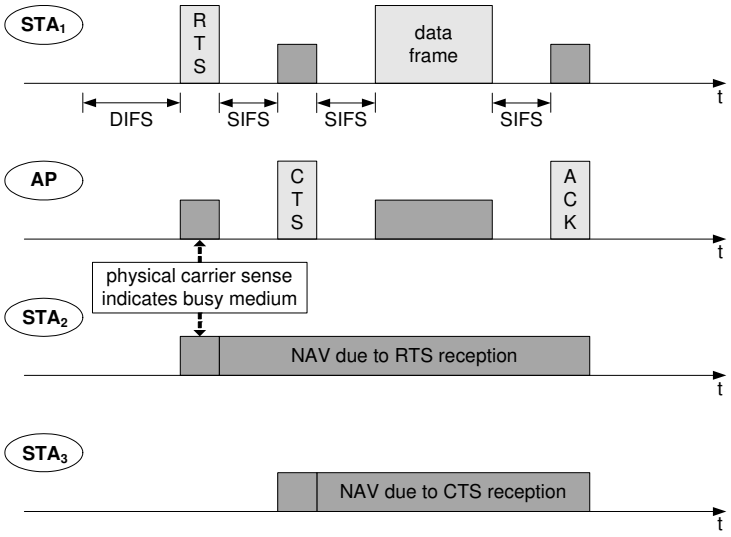


Figure 2.5: RTS/CTS for hidden terminal scenarios, modified after [1, Fig. 7.13, p. 177]

operating according to the MAC rules of the DCF. Out of the three inter-frame spaces, DIFS is the largest one. In contrast, *PIFS* is one slot time smaller than DIFS, but a slot larger than SIFS. PIFS is used by the AP to obtain prioritized medium access when the medium access rules of the PCF apply. Last, *SIFS* has the shortest duration and is used for frames that are immediate responses to others. An example of such frame is the immediate ACK that follows a data frame. As only the recipient of the data frame is allowed to transmit the ACK, this STA needs the highest priority for the channel access being realized with the smallest SIFS. We note that for 802.11 DSSS, HR/DSSS, ERP-OFDM, and DSSS-OFDM PHYs operating in the 2.4 GHz band, SIFS always has a duration of $10\,\mu s$.

RTS/CTS Extension It may happen that a STA is capable to receive transmissions from two other STAs which however themselves are not able detect any wireless signals from each other. This case is schematically shown in Figure 2.4. There, STA B is able to receive frames from A and C. In contrast, A is not able to detect any transmissions from C (and vice versa). Note that in this case, neither the physical nor the virtual carrier sense helps much, when for example A has an ongoing transmission to B, and C wants to initiate a transmission, too. In this case, C perceives the wireless channel to be idle, starts to send, and causes a collision of A's and its own transmission at receiver B. This scenario is known as the *hidden terminal problem*.

As a solution to this problem, 802.11 incorporates an optional means denoted as *request-to-send (RTS)/clear-to-send (CTS)* frame handshake. The basic operation is shown in Fig. 2.5 and works as follows. Suppose that STA_1 wants to transmit a data

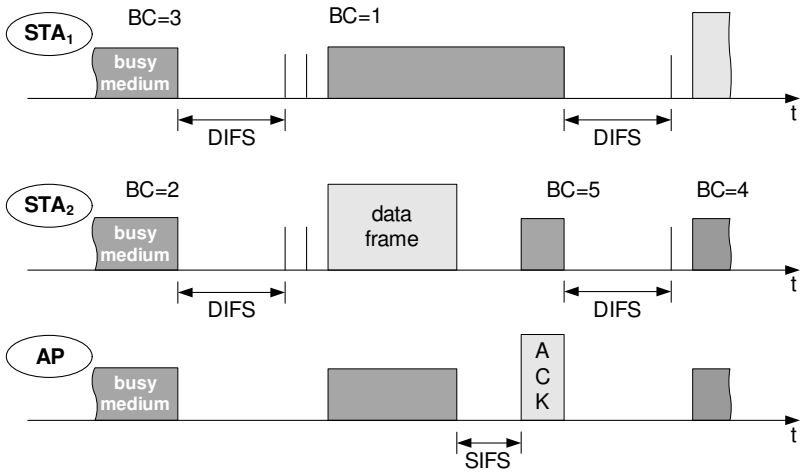


Figure 2.6: 802.11 backoff procedure, modified after [1, Figs. 7.11 and 7.12, pp. 175–176]

frame in the uplink to the AP. For this, it first sends out a small RTS frame, to which the AP as the intended receiver answers with a CTS frame. Both, RTS and CTS frames include the duration of the complete transmission sequence until the end of a potential ACK frame. With this, all STAs in the direct vicinity of *both* STA_1 and the AP are informed how long the wireless channel will be occupied and set their NAVs accordingly. Thus, these STAs are not able to initiate a transmission on their own in this time span.

2.6.2 Distributed Coordination Function

The distributed coordination function (DCF) specifies the basic medium access under which all WLAN STAs operate. Its basic principle is commonly known as *CSMA with collision avoidance (CSMA/CA)*.

According to the DCF rules, if a STA wants to initiate a transmission, it first applies the carrier sense mechanism that has to indicate an idle channel for DIFS. Once this condition has been satisfied, the STA may start its transmission. Otherwise, if the wireless medium has been busy, the STA conducts a procedure that is denoted as *backoff*, which is in essence a randomized waiting time. For this, the STA dices a random number of backoff slots, each with the size of a slot time, from a uniform distribution with the interval $[0, \text{CW}]$, whereby the *contention window (CW)* is initially equal to a predefined minimum value CW_{min}. The STA stores the diced number of slots in its backoff counter.

For an illustration of the backoff process, let us consider a scenario of an AP and two associated STAs as shown in Fig. 2.6. Let us assume that both STAs sensed a busy medium, while they have a data frame ready to be transmitted to the AP. As a result both STAs dice a backoff, whereby STA_1 has a backoff counter (BC) of three, while STA_2's BC is two. Now, once the channel becomes idle again, both STAs in the backoff stage first sense the channel for DIFS and afterwards divide the time into small slots. For each

elapsed slot time, the STAs decrease their backoff counters by one until either the counter reaches zero (STA_2) or the wireless channel becomes busy again (STA_1). In the first case, the STA immediately sends out the data frame. As the channel has become busy again, STA_1 has to resume its backoff, i.e., waits until the channel becomes idle again, then senses the channel for DIFS, continues to decrease its backoff counter, and so on.

If a transmission attempt of a data frame fails (i.e., an ACK has not been received), the sending STA also conducts a backoff prior a retransmission. For each retransmission, the STA thereby conducts a backoff for which it doubles its current $CW+1$ value as long as it stays below a given maximum CW_{max}. The intention behind this behavior is as follows. In WLAN BSSs with high load levels, one reason for several subsequent retransmissions are *collisions*. Such a collision may occur if two (or more) STAs have detected an idle medium for DIFS (and their remaining backoff slots), they are allowed to send and thus start their transmissions at the same point in time. Now, the intention of the growing CW values for the backoff is to avoid such potentially repeating collisions among STAs by simply increasing the range of their uniform distribution. This is the part that is known as the collision avoidance (CA). In total, for the basic data / ACK frame sequence, the 802.11 standard specifies an upper limit of seven consecutive transmission attempts per data frame. For the optional RTS/CTS feature, 802.11 allows seven consecutive trials for the transmission of a RTS[3] and four data transmission attempts. If these retry limits are reached, a STA drops the data frame and resets its CW to CW_{min}.

Finally, once a STA has been able to transmit a data frame, after receiving an ACK, it sets its CW back to CW_{min} and conducts a backoff procedure again (denoted as *post backoff*). In Fig. 2.6, this case is shown for STA_2 that conducts a backoff after the reception of the ACK frame. Such a post backoff is essentially a matter of fairness avoiding that a STA transmits packets back to back without enabling other members of the BSS to access the wireless channel.

2.6.3 Point Coordination Function

The point coordination function (PCF) is an optional extension that operates on top of the DCF rules. The intention behind the design of the PCF is to support delay-bounded traffic by enabling a timely and quasi-periodic delivery of data frames. The central component thereby is the *point coordinator (PC)*. It resides on the AP and regulates the medium access in a centralized fashion as a master by polling STAs for transmissions. As a basis, the PC divides the time into alternating *contention-free periods (CFPs)* and *contention periods (CPs)*, whereby the PCF is applied during the CFP and the DCF is used in the CP.

CFPs are scheduled at distinct TBTTs. To start a CFP, the PC tries to obtain the control about the BSS by accessing the wireless channel after an idle time of PIFS. Thus the PC has prioritized access compared to STAs following the DCF rules. Then, the PC initiates a beacon transmission in which it indicates the start of a CFP. All STAs

[3]Once a CTS has been received, the STA again has seven trials for an RTS, e.g., if the transmission of the data frame fails and the RTS/CTS procedure has to be repeated.

receiving this beacon set their NAV to the duration of the CFP, thus being not able to access the wireless channel in between on their own. After the transmission of the beacon, the PC waits for SIFS before conduction subsequent actions.

During the CFP, each STA is only allowed to conduct an uplink transmission on request. The PC maintains a list of STAs that should be prompted, being denoted as the *polling list*. The PC sends a special *contention-free poll (CF-poll) frame* at least once per CFP to each STA on this list. After receiving a CF-poll, the corresponding STA is allowed to initiate an uplink transmission after SIFS. In case of a successful reception, the PC replies after SIFS with a *contention-free acknowledgment (CF-ACK) frame*. In contrast, if the PC does not receive a reply to its poll, it gains control over the wireless channel again after PIFS.

The PC may send data frames in the downlink to the STAs, may poll STAs to allow an uplink transmission, and may acknowledge an uplink transmission. Polled STAs may transmit a frame in the uplink and may acknowledge a preceding frame in the downlink. For all of these actions, frames in one direction may be 'piggybacked', meaning for example, that a data frame to one STA is combined with a poll for an uplink transmission. Such piggybacking can even include a frame addressed to another STA. Finally, the CFP ends if the maximum specified duration is reached or the PC explicitly sends a *contention-free end (CF-end) frame*.

2.7 QoS Extensions

The two medium access functions discussed above have shown significant drawbacks in combination with data traffic that needs to fulfill tight QoS requirements regarding end-to-end delivery delay, frames losses, and minimum throughput. These issues have been extensively discussed in the literature, whereby we point the reader for a comprehensive survey to Ni et al. [19]. In the following, we summarize these aspects, then survey the solutions included in the 802.11 standard today (as amended by 802.11e in 2005), and finally describe the principles of one of these medium access schemes being primarily used in this thesis besides the basic DCF.

2.7.1 Issues with the Basic Medium Access

The basic WLAN medium access inside a STA only relies on one interface queue above MAC in which the data packets coming from the upper layer are stored before being processed and transmitted. These queues usually obey a first-in-first-out (FIFO) discipline such that packets are handled in the sequence of their arrival. This completely prohibits a differentiated handling of data packets belonging to different types of traffic streams. It can happen, for example, that time-bounded VoIP packets have to wait in the queue until packets with more relaxed delay constraints have been transmitted, thus imposing a significant, additional delay.

Further at MAC, the DCF with its backoff scheme handles all traffic types in the same fashion without a possibility of any prioritization. Specifically at higher load

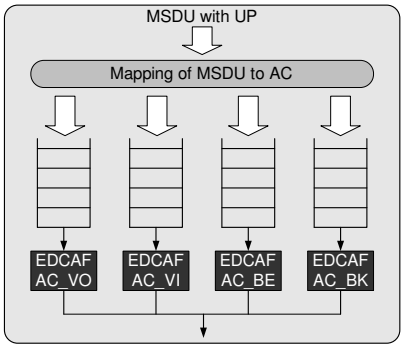


Figure 2.7: Priority queues and EDCAFs of an 802.11e STA, modified after [6, Fig. 9-19, p. 874]

levels in a WLAN BSS, this may lead to unpredictable delays for all types of traffic as a result of an increasing number of retransmissions together with repeating backoffs. Further, STAs with voice traffic compete against other STAs having delay tolerant traffic flows, although being essentially the first suffering from increasing delays and losses. In summary, the DCF provides fairness regarding the number of transmitted frames, but not regarding the amount of transmitted data.

Although the PCF was designed for a support of QoS-constrained traffic, it actually suffers from different aspects such that it has been rarely used in practice. First, the start of a CFP with the corresponding beacon frame may be delayed as the channel may still be occupied by a transmission from the preceding CP phase. Further, a polled STA may transmit a frame up to a maximum length of 2304 Bytes. In combination with low link data rates, this can significantly delay subsequent transmissions of other STAs. Thus, an exact timing of polled transmissions is hardly possible. Lastly, for an exact polling schedule, the PC may need information about the specific type of traffic and its characteristics, e.g., regarding packet inter-arrival times and minimum required throughput or tolerable loss rates.

2.7.2 Medium Access Enhancements

To circumvent the issues discussed above, the 802.11e amendment [20] introduced the *hybrid coordination function (HCF)* back in 2005 enhancing the legacy 802.11 medium access functions. HCF consists of the *HCF controlled channel access (HCCA)* and the *HCF enhanced distributed channel access (EDCA)*. The HCCA is a polling-based medium access function similar to the PCF. However, it allows the central instance at the AP, denoted as *hybrid coordinator (HC)*, to schedule polling phases not only in CFPs but also during CPs. Further, a polled STA may send multiple data frames in the uplink. For this, each poll includes a maximum duration for the uplink transmission(s) of each STA. Still, for an efficient operation, the HCCA may require quite sophisticated information for a well-matching polling schedule about each traffic stream (e.g., mean or peak values

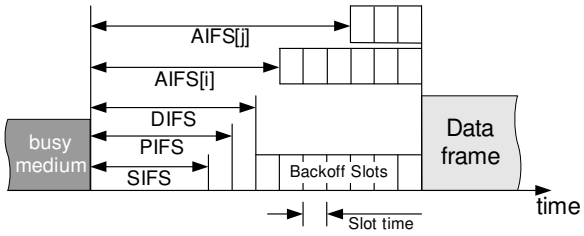


Figure 2.8: Inter-frame spaces with 802.11e EDCA, modified after [6, Fig. 9-3, p. 826]

of data traffic rates or frame sizes). Although 802.11e includes means for the signaling of such parameters on a traffic stream level between a STA and an AP, the algorithms to obtain such polling schedules are vendor-specific, opening a broad range for different approaches. For a survey about the scheduling discussion, we refer the reader to [21–24] and the references therein. As we do not want to make our work dependent on a specific type of scheduling algorithms, this thesis bases not on the HCCA but on the EDCA when considering QoS extensions, thereby being mainly driven by its simplicity.

2.7.3 Principles of EDCA

The EDCA follows in principle the CSMA/CA medium access, but extends the legacy DCF regarding several aspects. In order to allow a prioritized and separate handling of different traffic types, the EDCA basically utilizes the following three distinct means.

First, data traffic that arrives from the network layer is mapped to different *access categories (ACs)*. The mapping of the data traffic to these ACs essentially relies on the *DiffServ code point (DSCP)* [25], included in the *IP header* of each data frame (for IPv4: within the type of service field, for IPv6: within the traffic class field). The DSCPs are used for a classification of traffic and are further mapped to eight *user priorities (UPs)* given in IEEE 802.1D [26]. The 802.11e amendment finally specifies the mapping between the UPs and four different ACs—for Voice (*AC_VO*), Video (*AC_VI*), Best-Effort (*AC_BE*), and Background (*AC_BK*) traffic. Thereby, AC_VO has the highest and AC_BK the lowest priority.

Second, each of the four ACs is equipped with an own FIFO transmit queue. By this, traffic arriving at MAC level is buffered according to each priority level separately in one of the four queues as shown in Fig. 2.7. This prevents that the queuing time of high priority data traffic is influenced by queuing delays of lower ACs.

Third, 802.11e introduced an *EDCA function (EDCAF)* per AC. Each EDCAF contends separately for the medium access during a contention phase only for traffic of its AC. For an additional support of high-priority traffic, 802.11e introduced the possibility to initiate per contention phase multiple transmissions in a row, which are separated by SIFS. This is denoted as *transmission opportunity (TXOP)*, whereby the *TXOPLimit* determines the maximum duration of all the frames allowed to be transmitted in sequence.

Table 2.2: Default parameter set

AC	CW_{min}	CW_{max}	AIFSN	TXOPLimit	
				DSSS and HR/DSSS PHYs of 802.11/b	OFDM and ERP PHYs of 802.11a/g
AC_BK	aCWmin	aCWmax	7	0	0
AC_BE	aCWmin	aCWmax	3	0	0
AC_VI	$\frac{\text{aCWmin}+1}{2} - 1$	aCWmin	2	$6.016\,ms$	$3.008\,ms$
AC_VO	$\frac{\text{aCWmin}+1}{4} - 1$	$\frac{\text{aCWmin}+1}{2} - 1$	2	$3.264\,ms$	$1.504\,ms$

Extending the basic DCF, each EDCAF has an own contention parameter set. The default parameters as given by 802.11e are listed in Table 2.2 and consist of the upper and lower bounds of the contention window (CW_{min} and CW_{max}), own inter-frame spaces, and a *TXOPLimit* values per EDCAF. Note that the CW parameters depend on the actual WLAN PHY, whereby *aCWmin* and *aCWmax* are the standard CW limits as specified for the given PHY configuration. Further, the inter-frame spaces of the EDCA are denoted as *arbitration inter-frame spaces (AIFSs)* and are defined as follows

$$\text{AIFS[AC]} = \text{SIFS} + \text{AIFSN[AC]} \cdot t_{\text{slot}}, \tag{2.1}$$

whereby t_{slot} is the duration of a slot time and *AIFSN* specifies the number of slots for each AC. The lower the priority of an AC, the larger is usually its AIFS value. In other words, lower priorities have to conduct a longer carrier sense leading on average to a lower probability for a transmission initiation or a backoff decrement compared to the highest priority. Figure 2.8 exemplarily shows the relationship of AIFS values belonging to different EDCAFs.

Similar to the legacy DCF behavior, each EDCAF starts a backoff when either the medium is busy upon frame arrival from its transmit queue, an ACK timeout occurs while waiting for an ACK, or a frame was transmitted successfully and the corresponding ACK was received properly. A transmission failure (ACK timeout) leads to an increased contention window of an EDCAF according to the binary exponential backoff algorithm, which doubles *CW* as long as it does not exceed CW_{max}. Further, after a successful transmission, CW is reset to the minimum value for the corresponding AC.

With EDCA, an *internal collision* between EDCAFs happens if at least two of them have the right to initiate a transmission at the same point in time. This case is resolved such that the EDCAF with the highest priority gains access to the wireless medium, while the lower-prioritized EDCAF(s) perform(s) actions as if an external collision appeared.

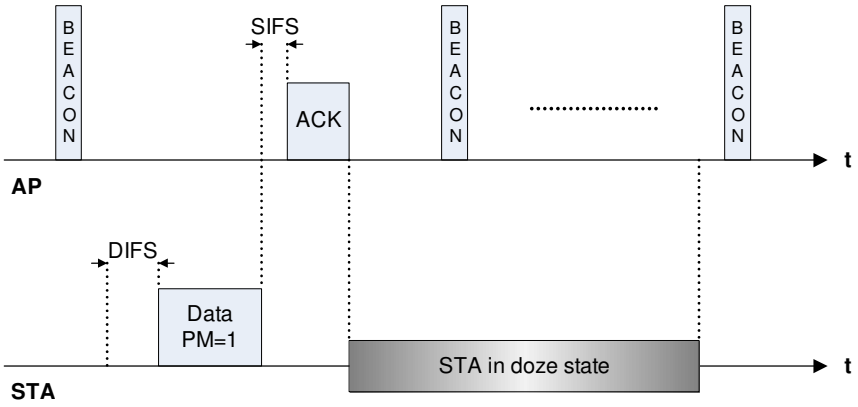


Figure 2.9: Initiating the power save mode: the STA sets the power management bit to one in an uplink transmission.

2.8 802.11 Power Management

802.11 technology is commonly applied in battery-powered devices such as notebooks, tablet computers, or smart phones. To economize the energy expenditures during the use of WLAN access, the 802.11 standard introduced mechanisms for a power management. In the following, we first describe the STA states relevant to power management and afterwards discuss related mechanisms.

2.8.1 Power Management States and Modi of a WLAN STA

The IEEE 802.11 power management defines two different power states for a STA: the *awake* and the *doze state* [6]. Communication between the AP and the STA can only occur in the awake state. In the doze state, the "STA is not able to transmit or receive and consumes very low power" [6].

To handle these two different STA states, the 802.11 standard distinguishes between two different STA modi: the *active mode (AM)* and the *power save (PS)* mode. In AM, a STA always has to be in the awake state, implying that it is permanently capable to transmit or to receive. In contrast, a STA in PS mode may change between the awake and the doze state to decrease its energy expenditures. Thus, to avoid any losses of data frames for a STA in PS mode, the AP provides specific mechanisms enabling to buffer any pending downlink data for the corresponding STA.

2.8.2 Initiating the PS Mode

Figure 2.9 illustrates the frame exchange used by all standard-compliant STAs to initiate the PS mode. The STA has to go through the CCA procedure and may then send an uplink data frame in which the so-called *power management (PM) bit* of the frame control field has been set to one [6]. The following ACK confirms not only the reception

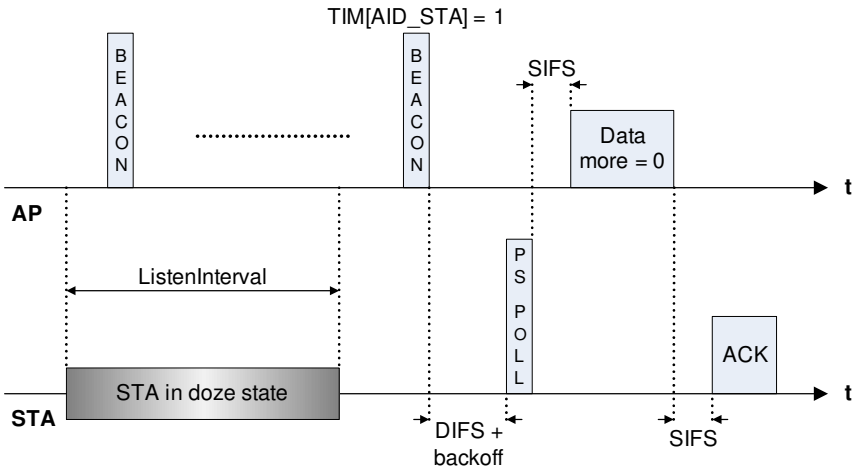


Figure 2.10: Wake-up for the reception of downlink traffic, modified after [6, Fig. 10-4, p. 986]

of the data frame, but also acknowledges that the AP has marked the STA being in PS mode. Then, the STA may change to doze state.

2.8.3 Periodic Wake-up Procedure

The AP buffers downlink data for a STA being in PS mode. Such pending traffic for a specific STA is indicated by a flag transmitted within the beacon frame. The element containing this flag is denoted as *traffic indication map (TIM)*. By means of the *timing synchronization function (TSF)* together with the knowledge of *target beacon transmission time (TBTT)*, the 'sleeping' STA knows about the points in time when these pseudo-regular beacons are expected to be transmitted. As illustrated in Fig. 2.10, the STA changes periodically a little earlier than TBTT into the awake state in order to receive a beacon. If the TIM indicates pending downlink traffic, the STA stays awake and triggers a downlink transmission with a so-called *PS poll frame*. The AP either directly replies with a data frame or just acknowledges the poll thereby delaying the actual data transmission. In case that the AP has more than one data frame for the STA, it sets the *More Data bit* in the data frame. Then, the STA initiates a further PS poll. This repeats until a downlink data frame is received in which the More Data bit is set to zero.

Note that this power management results on the one hand in long doze times of STAs in PS mode. However, on the other hand, it may lead to unacceptable delivery delays for data frames in the downlink direction as their presence is just announced in beacon frames, which are periodically transmitted only over large intervals being in practical networks usually around 100 ms. Even worse, the STA usually remains in the doze state for some contiguous beacon intervals further increasing the delay. Thus we can conclude that the legacy wakeup procedure is not suitable for highly delay-constrained traffic such as VoIP at all.

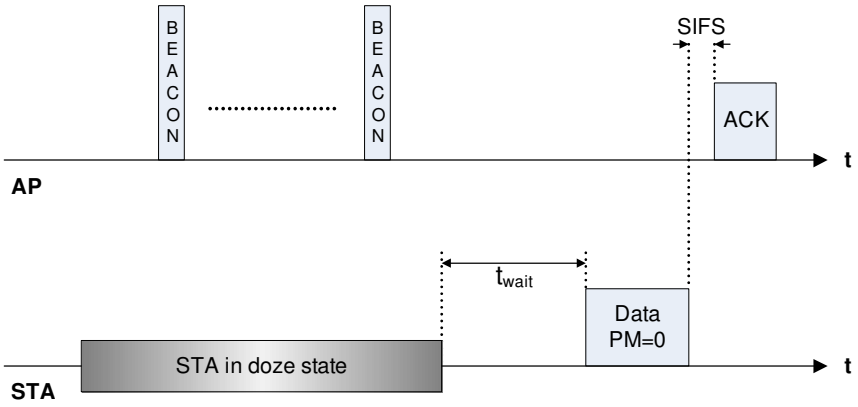


Figure 2.11: Terminating the PS mode: the STA sets the power management bit to zero in its uplink transmission

2.8.4 Immediate Wakeup Procedure: Terminating the PS Mode

A STA in PS mode is allowed to resume uplink communication at any time by switching from the doze back to the active state. Thus, such a STA may not only transmit uplink data but can further terminate its PS mode at any time returning back to normal, active mode as shown in Fig 2.11.

The detailed steps terminating the PS mode are as follows. First the STA has to sense the wireless channel for a given time (t_{wait}) after changing into awake state or has to successfully obtain a current setting for the network allocation vector. This may help that the station does not interfere with any ongoing transmission even if the physical CCA indicates an idle channel (hidden node problem). Afterwards, the STA competes for channel access and transmits an uplink data frame signaling its change from PS to AM by setting the PM bit of the frame header to zero. If the STA has no pending uplink data, it may in turn send a *null-data frame*. Such a null-data is a MAC-level data frame consisting only of the MAC headers, but as its name implies, without any data payload.

Chapter 3

Resource-aware Handovers

This chapter starts by surveying reports and forecasts about the trends regarding data traffic in WWANs, specifically regarding 3GPP cellular networks. Then we review existing strategies to shift traffic from such cellular networks to other wireless accesses comprising of small cell technologies such as femto and WLANs. The evolution regarding such a traffic offloading involves vertical handovers, where data traffic streams are migrated to small cell accesses. Thus, we introduce the definition of a vertical handover and describe related mechanisms. Then, we argue about the selected direction of handovers for this thesis by considering their objectives, related policies, and requirements for a realization. Finally, we consider mechanisms for traffic offloading, give an overview about recent frameworks from the IEEE and the 3GPP standardization bodies and argue in detail about the latest support for IEEE 802.11 networks regarding radio resource measurements and network management. Certain parts of our survey have been published before in [27–29].

3.1 The Capacity Crunch

The global trend towards a rapidly growing number of new devices such as smart phones or tablets together with emerging traffic-intensive applications such as video has led to an increasing amount of overall mobile data traffic in the last years. The term *mobile data traffic* thereby refers to all data traffic that is served by operators in their 3GPP cellular networks. Specifically network equipment vendors such as Cisco have been trying to measure and to predict the evolvement of this type of traffic thereby relying on several measurements of their customers (i.e., network operators) as well as the predictions of manifold analysts. Over the past years, Cisco has been publishing its annual forecast about the mobile traffic on the basis of five-year look-ahead time frames. These reports gained a considerable amount of attention as they have been continuously predicting a strong, exponential-like increase in the mobile data traffic. Unfortunately, old reports of Cisco are almost not available on the Internet anymore, however, we have been archiving

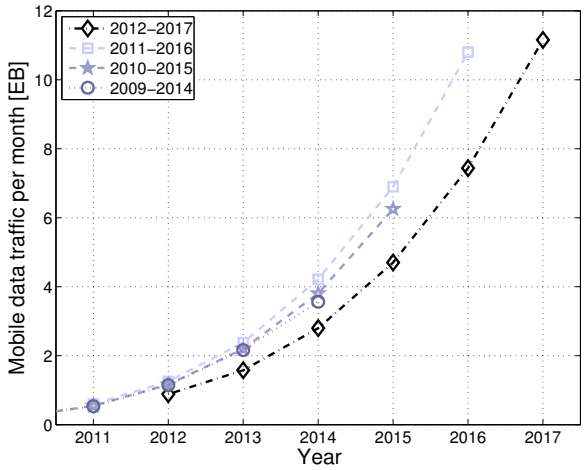


Figure 3.1: Cisco's predicted global mobile data traffic from 2009 to 2017; data points before 2011 are very close to each other and thus excluded for a better visibility.

these reports starting in 2009, which allows us to conduct a brief comparison of the reports being published between the years 2009 and 2013 [30–34].

Figure 3.1 shows the forecasted average of global mobile data traffic per month for the four different reports from 2010 to 2013[1]. Note that the mobile traffic is given in exabytes, whereby one exabyte (EB) equals $1,000,000$ terabytes. While the precise measurement and estimation methodology including the underlying models for these reports have not been published in detail, a brought spectrum of people ranging from analysts to researchers use these results as an argument for an ever increasing mobile data traffic. Although the reports published from 2010 to 2012 indicate consistently a strong manifold increase of the global mobile data traffic, estimating a quasi exponential growth, they do not give any hints about an upper bound further. Some critical voices appeared on the web from analysts, e.g., at [35], questioning Cisco's forecasts by arguing that the actual traffic increase essentially appears to be much slower.

Interestingly, even the latest Cisco forecast for the period of 2012 to 2017 indicates a much slower increase in mobile traffic compared to the previous reports. This is also visible in Figure 3.1, where the curve of the latest forecast shows much lower values for 2012 to 2016 compared to the graphs of the previous reports. Cisco itself gives three reasons for this behavior [34]. First, they argue that limited flat rate models for mobile data traffic, so-called capped or "tiered" data plans, in the Western European region have dramatically slowed down the traffic increase. The principle behind these limited flat rate plans works as follows: once the mobile user reaches its monthly traffic limit, he is throttled down to very low traffic rates. Second, Cisco reports on a global level a

[1]The results published in 2009 regarding the overall mobile data traffic do not differ much from the 2010 report, thus they were not included in Fig. 3.1.

smaller growth regarding the number of laptop devices accessing cellular networks. In the past, this class of devices was shown to contribute to the mobile traffic considerably. Third, it seems that a high fraction of mobile traffic has been shifted to other wireless access technologies such as WLAN hotspots. Some analysts argue that the impact of the latter aspect has been constantly undervalued by Cisco over the last years [35], such that even the current predictions still show exalted results [36].

Considering further scientific studies regarding the actual appearing traffic in cellular networks, it becomes clear that the distribution of the data traffic load is essentially skewed among both, wireless cells as well as end users [37, 38]. Based on traces from a cellular network for one week back in 2007, Paul et al. [37] reported that on the one hand about 90 percent of the mobile data traffic was created by less than 10 percent of the end users. On the other hand, 50 to 60 percent of the overall data traffic was transported by just 10 percent of the wireless access cells. Note that this study was conducted shortly before the smart-phone wave started. In a more recent study from 2010, Jin et al. [38] identified that about 30 percent of the load was driven by one percent of end users, while other 20 percent of the overall traffic was generated by just two additional percent of end users. Further, the authors analyzed the behavior of users with more than one gigabyte (GB) data traffic per month being denoted as "heavy users". Their findings showed that video and audio streaming applications were the dominant reason regarding the high traffic consumption for about 40 percent of heavy users. Furthermore, these heavy users were reported to cumulate just at a few places, whereby they most often did not seem to change their location [38].

In summary, even if one may be generally cautious regarding forecasts claiming a trend of rapid, ever-increasing mobile traffic, from the works of Paul et al. [37] and Jin et al. [38] it becomes clear that at least temporarily a certain fraction of cellular network cells essentially has to deal with a strong load from certain users so that the cellular capacity may be exhausted and congestion occurs in this specific region. The following section discusses solutions that have been presented in the literature so far for this specific problem.

3.2 Offloading Traffic from Cellular Networks

To deal with the increasing load in their cellular networks, operators apply a broad range of solutions. Namely an increase of cellular network capacity, capped data plans, and a shift of traffic to other wireless access technologies have been briefly described in the literature [39, 40], whereby the latter has been argued to be probably the most suitable solution. The reasons for this are as follows. As a bare increase of the overall cellular capacity by a simple expansion of the networks comes up with tremendous financial efforts, it remains more than questionable whether just a continuous and strong provisioning of additional resources can be the complete solution for cellular operators. Lee et al. argue that such a strategy will likely not pay out especially if the revenue of mobile data operators does not scale with the amount of transported data consumed by its customers [40]. As a result, operators (specifically in Western Europe) tend more and more not to offer

unlimited data flat rates, but instead provide "tiered" data plans which come with a fixed limit of monthly traffic volume, throttling down their customers to low data rates once their monthly limit is exhausted [34]. However, this only caps the overall traffic for each customer per month. Thus, it still allows a strong spatial accumulation of high traffic demands by a high number of customers within the cellular network leading to congested cells on a temporary basis. To deal with all these circumstances, the most suitable solution seems to be a shift of traffic to 'small-cell' wireless access technologies such as WLANs or femto cells. This shift of traffic is denoted in the literature as *offloading* [39, 40]. Today, the most prevalent case thereby consists of the movement of traffic to WLAN access cells, whereby manifold reasons exist for the usage of WLANs as the most dominant offloading technology. Its popularity over the last decade leads to the situation that WLAN is available in the large majority of end-user devices today. As such it is broadly present in a wide range of application fields ranging from home deployments over single public hotspots to large enterprise wireless access networks. In addition, WLANs are usually connected to the Internet via own wired networks not touching the cellular infrastructure, thus making the offloading very advantageous from the operator's perspective.

Let us now have a closer look at the different offloading strategies that are either already applied today or envisioned for a use in the future. We categorize the existing strategies according to the point in time *when* data traffic is offloaded to a second technology. There, different approaches range from an immediate offloading to time-shifted strategies. In a nutshell, *the immediate heuristics* apply WLAN access just whenever it is available. That is, once a mobile customer is able to connect his device to a WLAN hotspot, he uses this wireless connection for the transmission of all of his new data sessions. If the WLAN connectivity becomes unavailable, e.g., because of mobility, depending on the management capabilities of the device, the data sessions may be re-initiated within the cellular networks from application level or may simply break. Note that this immediate offloading is the de-facto solution today, decided and applied by mobile customers as they naturally prefer to use WLAN access when available for their data sessions. Besides being motivated by the popularity and the availability of WLANs, this user behavior is also strongly influenced by the common tiered data plans offered by cellular operators. In order to balance the amount of their monthly data volume carefully, end users usually prefer to have their data sessions in their accessible WLANs, e.g., at home or in their offices.

Several studies have been analyzing the amount of traffic that can be offloaded by immediate heuristics, thereby considering diverse user behaviors with different traffic demands essentially leading to strongly diverging results. Although one should be cautious about an exact comparison of previous works because of different assumptions as well as user and traffic models, we give an overview about the contributions so far by highlighting the different underlying scenarios. Lee et al. [40] reported that 65 percent of their mobile traffic could be served by WLAN hotspots with such a simple strategy as the considered mobile customers on average stayed 70 percent of their time in the coverage area of WLAN cells. Their offloading results based on simulations considering real WLAN connectivity traces of 100 iPhone users collected in different Korean cities

over a period of two and a half weeks in 2010. Further, also on the basis of simulations with WLAN connectivity traces, Balasubramanian et al. [39] analyzed the amount of traffic that can be shifted to WLAN cells for vehicular movement patterns. The authors gathered traces from three different cities in the United States, considering 20 busses in the first setting as well as private vehicles in the two remaining ones. The measurement duration varied from 12 down to 3 days. For such a vehicular environment, their results showed that up to 23 percent of the considered mobile data traffic can be offloaded by following simple heuristics. For eight weeks in the beginning of 2012, Liu and Striegel [41] conducted an on-campus measurement campaign with a pool of 131 students equipped with smart phones and WLAN access in their dormitories. Their measurements indicated that rather a maximum amount of just 33 percent of the mobile traffic may be offloaded in the considered environment.

The second group of strategies which we denote as *time-shifted offloading* varies the point in time when traffic of an end-user device is served. We further differentiate between two basic approaches in this category. When WLAN access is available, data may be loaded far before an end user actually demands a specific content. This is also known as "prefetching" [42]. An example for such a prefetching strategy is a download of emails or newspaper content to an end device via a WLAN network at user's home, such that the data can be accessed without using cellular networks while the user will be on the move later. In the opposite direction, traffic demands of an end user may be delayed in time, either until WLAN connectivity becomes available or a given deadline expires that is usually application-dependent, i.e., how long the end user is willing to wait for his content. This is also known as "delayed offloading" [40].

For their vehicular scenarios, Balasubramanian et al. [39] showed by trace-based simulations a strong increase of the amount of data that can be shifted to WLANs for the case of delayed offloading. For a 60 second deadline, the authors show that on average 45 percent of the considered traffic can be served by WLANs (compared to just 23 percent for immediate offloading). In contrast, Lee et al. [40] showed differing improvements of delayed over immediate offloading for their traces obtained from 100 iPhone users in everyday life. While deadlines for a delay of 100 seconds show only a small increase in traffic shifted to WLANs, for considerable increases of 21 to 29 percent, the authors suggest to use large deadlines like an hour and even above. In summary, similar to the immediate heuristics, also for the delayed offloading the amount of shifted traffic strongly depends on the user behavior regarding employed applications and their resulting traffic demands as well the mobility patterns. Tackling these aspects, Ristanovic et al. [43] presented one approach, "HotZone", that bases on a mapping of user mobility and location information, thus allowing to predict possibilities for delayed WLAN offloading. Further, Balasubramanian et al. [39] proposed an approach, "Wiffler", that estimates for a device on the basis of measures from the recent history, how many WLAN APs it will discover in a given time frame and which throughput it can expect in each hotspot.

For the sake of completeness, we briefly discuss time-shifted offloading strategies where the mobile traffic is not moved to WLAN hotspots, thus being not within the primary focus of this thesis. Instead of offloading traffic to WLAN hotspots, this type

of strategies relies on so-called *opportunistic communications* [44], which is essentially a form of delay-tolerant, direct communication between mobile end devices. In other words, when mobile devices are temporarily in the vicinity of each other, direct communication is enabled via short-range wireless technologies such as Bluetooth or WLAN ad hoc modes. Offloading by opportunistic communications works as follows [44, 45]. A certain subset of mobile devices initially downloads the content of interest via the cellular networks and afterwards relays it to others if devices meet each other as a result of mobility. Han et al. [44] focussed on strategies for the selection of first, suitable candidates downloading the content such that the overall traffic in the cellular network is minimized. For direct device communication, Baier et al. [45] based their strategies on estimates regarding locations and speed of mobile devices. Further, Ristanovic et al. [43] presented a solution, "MixZone", in which the cellular operator coordinates opportunistic communication among mobile devices being in the coverage of each other.

3.3 Heterogeneous Handovers: A Tool for Offloading

Today, devices usually have connection possibilities to different wireless access technologies from the WLAN and WWAN area. As a result, such a device can perform wireless access over either of the different network types. If a device further has different NICs available, in addition, even a parallel operation of technologies is possible. Previous works about traffic offloading such as [39, 40] base on the assumption that data traffic sessions may be shifted to WLANs and further can be resumed by WWANs once WLAN access becomes unavailable. While delay-tolerant applications such as email may accept interruption times due to a switch of the network access up to several tens of seconds, specifically time-constrained services such as VoIP or Video streaming have tight requirements regarding a smooth migration. We understand an offloading to be *seamless* if the interruption of a service on application level due to the switching of the data traffic is so small that it is still acceptable for the end user.

3.3.1 Definition of a Heterogeneous Handover

We refer to the process of shifting on-going traffic between such two heterogeneous networks as *vertical or heterogeneous handover*. In contrast, for the sake of completeness, a handover between two cells of the same technology, e.g., two WLAN hotspots, is denoted as *homogeneous handover*.

For a further distinction regarding handovers, we follow mainly the common definitions summarized by Manner et al. [46]. A handover can be done either reactively (*break-before-make*) or proactively (*make-before-break*) referring to whether the involved steps of the handover are conducted before or after the device is loosing connectivity of its original wireless access link. Two other degrees of freedom in classifying a handover are given by considering which entity is initiating and which one is controlling the handover process. A handover may be either initiated, i.e., triggered, by the network or the mobile-device side. We refer to this as *network-initiated* vs. *mobile-initiated hand-*

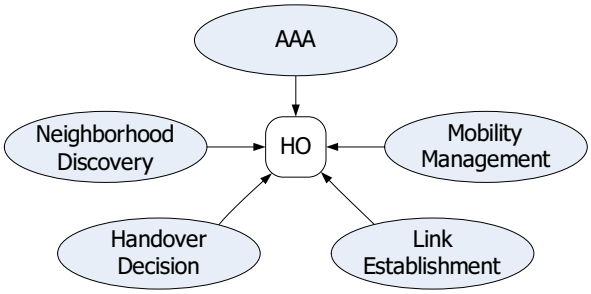


Figure 3.2: Classes of mechanisms affecting the handover process

over. Further, the actual handover process may be entirely under the control of the device or it may be controlled by a network entity of the provider. We refer to this as the *mobile-controlled* vs. the *network-controlled handover*. Although controlled by one of the entities, the handover processes may be assisted by the counterpart, leading to *mobile-controlled, network-assisted* vs. *network-controlled, mobile assisted* handovers. In addition to this distinction, the device may have also several link-layer connections active in parallel during the handover process to different points of attachments. We refer to a *point of attachment* as a piece of network equipment with which the device has a link layer connection. Having multiple such connections, a device may receive packets from several points of attachment simultaneously which avoids packet loss and is hence denoted as *soft handover*. In contrast, a *hard handover* first releases or interrupts its existing link layer connection, before setting up a link with another point of attachment.

Note that the handover process itself does not only consist of the bare switching between wireless links but further includes a couple of distinct mechanisms. In the following, we give a brief overview by grouping these mechanisms regarding their functionality.

3.3.2 Taxonomy of Handover Mechanisms

Figure 3.2 depicts the basic categories of a handover process that includes mechanisms regarding neighborhood network discovery, handover decision, link (re-)establishment, mobility management, as well as authentication, authorization, and accounting (AAA). Also referred to as handover phases, these mechanisms are usually conducted in a sequential order, while some may even allow for a parallel operation.

Within the *neighborhood discovery* phase, the end user device becomes aware of available radio cells serving as a potential handover target. To obtain such information, mechanisms characterized by the underlying technologies as well as technology-independent methods are applicable. The former typically involve so-called scanning procedures in which the device either passively listens on the wireless channel for possible communication partners or actively probes potential candidate cells. The latter may involve an (external) information service which can be queried for neighborhood information depending on the current position of the device. For this, the involved signaling may

be specific to the underlying wireless technology (e.g., IEEE 802.11k MAC level neighborhood reports, cf. Sec 3.7.1) or technology independent (IEEE 802.21's information service, cf. Sec. 3.6.3). In principle, the neighbor discovery may be conducted on a regular basis, or may be event-driven, i.e., triggered by the decision that a handover is imminent.

The purpose of the *handover decision* is to determine *when* to conduct a handover for *which user* to *which target radio cell*. The latter issue is known in the literature as the *network selection problem*. All three aspects of the handover decision broadly depend on the intended objectives of a vertical handover. While we are highlighting these different objectives together with their state of the art about handover policies in the following section, let us briefly give an overview that gives a notion about the space of objectives. For example, a handover decision for a device that is about to move out of the coverage of a wireless cell is usually quite different from cases where an operator of a network re-arranges the associations of devices to wireless cells. The first case tackles the QoS for the end user perceiving already a degradation of his wireless link. In contrast, the second direction rather aims to maximize the number of devices in wireless cells from the perspective of the network operator. Such differences in the objectives of the handovers lead to different requirements from the end-user as well the operators' perspective thus involving diverging decision criteria criteria and handover mechanisms.

For the preparation of a handover, the device needs to set-up a wireless connection with the new point of attachment of the selected network. We refer to this step as the *link establishment* phase. This step involves a signaling between the device and the point of attachment which can, depending on the specific wireless technology, even extend beyond the mere wireless link into the wired part of a network operator. In summary, a handover results in a change of the network topology as the mobile accesses the network via the newly established link with the selected point of attachment. Depending on the actual handover scenario, a device may stay within the same administrative domain or move between two different domains—the latter is denoted as *inter-domain* handover [47]. As a domain, we understand here the network that is under control of one administrative instance such as a provider or a company. Since an inter-domain handover implies also a change of the IP subnet, it requires *mobility management schemes* [47]. For the sake of completeness, we note that depending on the applied technology and the size of the network in one domain, also intra-domain handover may require a mobility management. The mobility management schemes migrate the traffic of a device from one to another network. Possible schemes range from pure layer-2 mechanisms just for intra-domain handover, to layer-3 based mechanisms, over transport layer solutions, up to approaches on application level for both intra- and inter-domain handover. All these approaches vary broadly in terms of the involved signaling cost and the resulting delays (and delay jitter) of arriving packets at the mobile.

In addition to previously discussed, access-technology-related aspects, a handover may also require a (re-)authentication of the mobile and the target access network, authorization for the usage of network resources, and accounting for costs. Included functions are usually specific to the access technology, whereby current standardization bodies are tackling the interworking of different technologies also from the AAA perspective.

3.4 Objectives and Policies for Heterogeneous Handovers

We start with a survey of different objectives for vertical handovers and motivate our selected scope on resource-aware objectives afterwards. In this context, we discuss existing work on policies and further focus on decision criteria, related technological parameters, and handover decision schemes.

3.4.1 Overview of Typical Objectives

A vertical handover basically involves different parties, namely the end-user with his communication device and network providers or owners of wireless access networks. Each of these parties may have own, possibly contradicting objectives regarding handovers. While in general, we can think about infinitely many optimization goals for handovers, we briefly discuss the most common directions for each of the two different parties in the following. Fig. 3.3 highlights the directions for user-centric and operator-centric objectives of a handover.

User-centric Objectives

People today are more and more used to connect to the Internet anywhere and anytime. While being on the move, users need handovers for a support of a *service continuity* when they leave the coverage of a wireless access cell. Besides mobility, an important role for an end user plays the quality of experience (QoE) with which he perceives a certain service of an application. For example, during the streaming of a video clip or a running VoIP call, end users prefer a smooth and continuous progress of the service. Users aim to avoid perceivable interruptions or hanging connections and varying qualities regarding image resolutions and sound. These aspects are governed by the QoS in terms of throughput, end-to-end delay, and losses that a wireless network is able to deliver. Thus, end users usually favor to *maximize the QoE* of an application service that directly translates for streaming or real-time applications such as video and VoIP into a *maximization of the QoS*. If a wireless access network is not able to deliver the desired QoS level, e.g., in terms of obtainable throughput, users may aim to improve the situation by switching to another wireless technology. The need for such handovers may emerge from natural limitations of the involved wireless technology or temporal shortcomings—e.g., as result of a varying quality of the wireless link, a user may wish to switch to a better network.

Further, an end user himself may favor certain networks because of his personal preferences. Other objectives may consider the energy consumption of his device, evolving monetary costs and business plans with network operators. The *energy consumption* plays a critical role for end users as their devices have become more and more battery-operated over the last years. Common, battery-operated devices today are notebooks and handhelds such as smart phones or tablets. An important aspect for an end user is how his device is draining energy from its battery for a connectivity to the Internet. Since the energy consumption for a transport of data via wireless networks varies

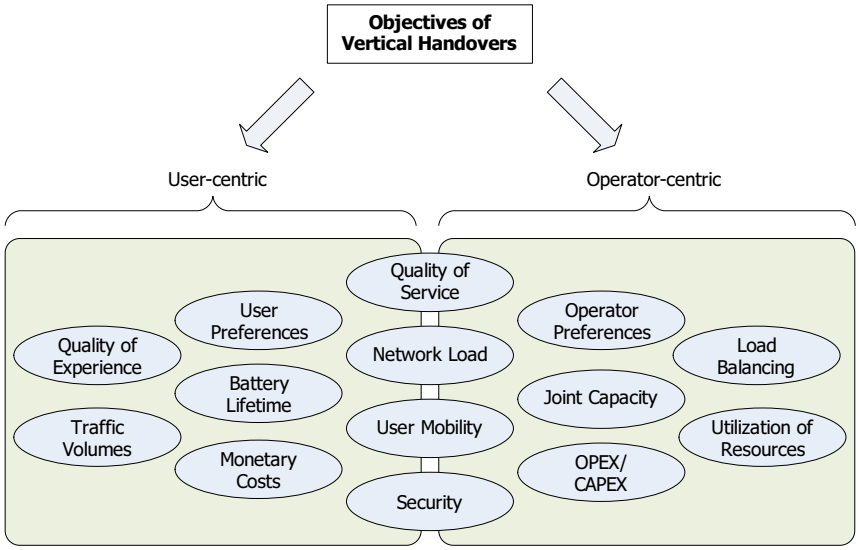


Figure 3.3: User- and operator-centric objectives for vertical handovers

greatly among heterogeneous technologies, a shift of the data traffic towards a more energy-efficient network may be plus.

Reducing monetary costs for the usage of wireless accesses can be another important objective for end users. While wireless accesses such as users' private WLANs as well as some public hotspots in cafes or shopping malls come for free, connectivity to other networks may impose significant costs. While at some exclusive spots, e.g., in hotels or at conferences, only fee-based WLAN access may be available, other wireless networks such as cellular accesses or even some public hotspots require a subscription plan with a dedicated usage model for data traffic. The most common plans today include limited flat-rates on the basis of a maximum data traffic volume being usually in the order of some hundreds of megabytes (MBs). A user exhausting his monthly traffic volume is usually throttled in his data throughput, just degrading the QoS of the cellular access rapidly. To *balance* his *monetary costs and monthly data traffic volumes* an end user may aim to carefully balance the access to these wireless networks.

This whole area of user-centric objectives is known in the literature also under the umbrella term *always best connected (ABC)* [48]. ABC aims to find in a given situation the best wireless access for an end device with a given traffic demand. We shall note that the term *always best connected* is thereby rather fuzzy as it does not describe well what is understood as the 'best connectivity'.

Operator-centric Objectives

Driven by their economical interests, operators of cellular networks usually aim to maximize the revenue of their networks. In this context, operators are faced with two different

types of costs for their networks, which are known as *capital (CAPEX)* as well as *operating expenditures (OPEX)*. CAPEX thereby includes all the costs for purchasing and installing additional equipment, e.g., to obtain a higher network capacity, while OPEX consists of the expenditures for "operation and management" of the cellular network [49]. Among others, OPEX also includes the costs for the energy consumption of the network equipment. In a nutshell, minimizing CAPEX and OPEX plays an important role for a cellular operator at different stages in the context of vertical handover objectives.

Operators usually aim to *balance the load* among available network resources and target to maximize the utilization of these resources, while still enabling mobility and QoS support for the end users. Vertical handovers specifically impose a strong gain as they enable to jointly use the capacity of other heterogeneous accesses. This is known as "capacity gain" that specifically evolves for complementary accesses such as WLAN and UMTS [50]. It is somehow obvious that the capacity gain increases with the number of heterogeneous accesses that do not interfere with each other. As a cellular network comprises not only of a wireless access but also of a wired backplane, an operator balances the load by considering all his network parts. Further, heterogeneous accesses may be even connected to the Internet via separate wired networks that are also not owned by the cellular operator. Thus, the costs for operation and management of the external wired and wireless networks are not included in operator's OPEX. Accordingly, shifting traffic to such accesses may essentially reduce the load for a cellular operator not only in the wireless but also in the wired part of his cellular network. The additional capacity of the heterogeneous access thereby comes essentially for free, i.e., without any additional CAPEX for the operator.

By means of vertical handovers, operators may further *improve the utilization of wireless resources*. Under an improvement of the utilization we understand to serve more end users with the same amount of resources, while still upholding a given QoS level for each user. Thus, an improved utilization may further reduce CAPEX and OPEX costs. The objectives towards a utilization improvement base essentially on different gains resulting out of the heterogeneity of networks. Besides the capacity gain, heterogeneity in wireless accesses as shown by Wang et al. [51] brings also other gains, which the authors denote as "multinetwork diversity", "multiuser diversity", and "spatial multiplexing gain". They stem from the fact that end devices may perceive different load and link conditions in each of the access networks. Utilizing this by allocating end devices to networks on the basis of their link conditions improves the utilization of wireless resources and is denoted as the "multinetwork diversity gain". Further reallocating the device with the strongest differences in link conditions has been thereby shown to be advantageous and is referred to as the "multiuser diversity gain". Lastly, the "spatial multiplexing gain" results from the fact that an end device has the option to use an alternative link if the perceived QoS in one wireless access cell has been suboptimal, e.g., because of a high load. In turn, this leads to improvements not only for the single device, but also for all other remaining participants in the original network.

Selected Scope of Objectives

In this thesis, we primarily focus on network-centric objectives for vertical handovers. Thereby, we go beyond the rather simple capacity gains which appear if one adds WLAN hotspots to a WWAN. More specifically, we aim to improve in WLANs how devices jointly use wireless resources in time, frequency, and space there. For this, we tackle a stronger utilization of these resources in WLANs, which may allow in turn, e.g., to maximize the number of end devices that can be offloaded from cellular networks.

We emphasize at this stage that we do not pursue a single, predefined set of objectives regarding vertical handovers. Instead, we target all objectives that, at least partially, aim to improve the utilization of WLAN resources. Accordingly, we denote this class of objectives as *resource-aware*. How to support such resource-aware objectives in WLANs in the context of vertical handovers is further detailed in the subsequent sections.

3.4.2 Resource-Aware Policies

We understand a *handover policy* to be a set of pre-defined rules formally describing a 'code of behavior' towards different and possibly contradicting objectives. The set of behavioral rules thereby specifies how the involved entities of a handover ranging from a device to the access networks shall act in a certain situation. We refer to any policy including a resource-aware objective as a *resource-aware policy*. While this class may include a variety of policies, we do not focus on specific flavors. Instead, we aim at a support and a realization of this class of policies in real-life networks.

Resource-aware policies exploiting the different gains by handovers between the wireless access networks have been discussed in the literature under different umbrella terms such as *traffic steering* [52, 53], *common radio resource management (CRRM)* [54, 55], as well as "operator motivated" vertical handovers [56]. From our perspective regarding offloading, Taha et al. [56] argue in the most comprehensive way about the policies and their requirements for operator-centric handovers.

According to our definition of a policy, it specifies a set of predefined rules for a resource-aware handover. Taha et al. [56] separate these steps into four phases, which the authors denote as "trigger", "identification", "selection" and "migration" phase. The comprehensive survey by Márquez-Barja [57] classifies other existing works into the "handover information gathering", "handover decision", and "handover execution phase". In essence, they are similar to Taha's identification, selection, and migration phase. As motivated above, we stick to the definitions of Taha et al. In their *trigger phase*, a network detects that handovers are needed for a reorganization of device associations to wireless accesses. This may depend, e.g., on the load level of one network part or new incoming connection requests. Then, the network analyzes in the *identification phase* which devices could be potentially conducting a handover. For this, the authors point out that such a choice should consider not only QoS per traffic stream and QoE levels of the end users, but must also take into account both wireless accesses, i.e., the status of the originator as well as the target cell, and applicable options of end devices. These applicable options per device include the available NICs on the end-device, an

observation of available networks for each device at its position, the number of previous handovers that the device had to conduct already and a consideration of the traffic streams regarding their suitability for handovers, e.g., in terms of acceptable interruption times or packet losses. In the approach of Taha et al., the status of the networks does not only include the load level, but further introduces the notion of "releasable bandwidth", which they further relate to the required signaling overhead per handover. This enables to balance whether a small number of users each with a significant load share or a high number of users with only a small load contribution should be considered for a handover. In a next step, the *selection phase* identifies the specific set of devices for a handover upon the information of the identification phase. This decision thereby bases on the objectives under which the reorganization of the heterogeneous network takes place. Finally, the actual handover together with all its signaling between networks and involved devices is done in the migration phase.

3.4.3 Requirements for a Realization of Policies

Realizing a handover policy with the different phases discussed above imposes several issues. First, criteria for handover decisions are required to reflect a given set of possibly contradicting objectives. In essence, the construction of such proper criteria usually relies on technological parameters. Usually, multiple of these criteria are a priori selected for an observation, such that in a next step a decision model can combine them in a predefined way. Such models are denoted in the literature sometimes also as *decision schemes* or *decision algorithms*.

Decision Criteria

Decision criteria in use for vertical handovers have been described in a myriad of work. The surveys [57–59] give an extensive overview about criteria being used for handover decisions. Márquez-Barja et al. [57] group the criteria according to the perspective that they reflect. The different groups essentially consist of network-, device-, user-, and handover-related criteria. While the interested reader is referred to reference [57] for a comprehensive list of criteria, we briefly summarize them by shortly discussing each of their levels. *Network-level criteria* range from link quality measures such as received signal strength indicator (RSSI), number of retransmissions, packet and bit error rates to metrics reflecting available bandwidth, aggregated network throughput, and load levels. On *device level*, criteria include available NICs, constraints for battery operation modes, location information, as well as the degree of mobility. User preferences about preferable networks and monetary costs for the wireless access give the *user-related* perspective, while *handover-related criteria* include the number of (previous) handovers, their success probability as well as their induced delays.

For our resource-aware policies, the challenging part is to reflect the underlying technological behavior that impacts how wireless resources are treated by each individual device. While mutual information about this behavior can be included in a large set of technological parameters, these parameters must be *available*. By availability we mean that

measures of technological parameters are accessible on a network component that computes the decision criteria. Basically, we can divide these parameters into three categories according to an accessibility at different network components. For a brief discussion, let us assume that we aim to gather the parameters on a selected piece of network equipment. A first subset of technological parameters may be directly measurable and accessible there, e.g., on a WLAN AP of an access network. In contrast, a second subset of parameters may instead not be directly available because it is measurable only at a disjoint piece of network equipment, e.g., on an end-user device. Lastly, the third subset consists of 'hidden' parameters which are only used inside internal modules of a network equipment. Typical representatives of the latter category are parameters that belong to closed, vendor-specific algorithms of a network equipment and are not exposed via typical management interfaces. The second category, where parameters are available only on disjoint network components, effectively calls for an appropriate orchestration of parameter measurements as well as their signaling between involved network components. We later discuss standard amendments such as 802.11k/v from the WLAN arena that tackle these aspects of the second category and show that this issue is not that critical although still an appropriate selection among the standardized measurement parameters has to be made.

In contrast, a problem appears in the context of the third category, where technological parameters are not directly accessible in either of the network components. If these parameters are critical with respect to a realization of certain decision criteria, they also become critical regarding an implementation and realization of the specific policy.

Handover Decision schemes

Several surveys tried to summarize and structure existing work on handover decision schemes [58–60]. Again, depending on the objective(s) of the corresponding handover strategies, a whole bunch of criteria may have to be taken into account. Thus, the question arises how to derive decisions even on the basis of a large pool of criteria. Over the last ten years, a variety of mathematical models have been presented in the literature to deal with this problem. The recent tutorial of Wang and Kuo [60] presented a comprehensive, in-depth survey about existing mathematical approaches ranging from "utility theory", "multiple-attribute decision making", "fuzzy logic", "game theory", and "combinatorial optimization", to the area of "Markov chains".

Kassar et al. [58] as well as Yan et al. [59] qualitatively compared existing approaches not only regarding the mathematical decision models but also with respect to the selected decision criteria. Kassar et al. classified existing work into "decision function-based", "user-centric", "multiple attribute decision", "fuzzy logic and neural networks based", and "context-aware" approaches. Decision function schemes are based upon cost functions as decision models that consider the network view, the user basis, or a combination of both. This class of approaches generally aims to identify either for each user or for each traffic stream the network with the smallest value of the cost function. User-centric schemes mainly base upon the perspective of the end user, e.g., regarding the obtainable QoS, resulting costs from the user perspective, and user satisfaction levels. Multiple attribute decision schemes take into account not only different decision criteria but also

sometimes even conflicting objectives. Utilizing multi-criteria decisions with fuzzy logic or neural networks based schemes handle the case that only inaccurate or noisy data for the criteria may be available. Lastly, the context-aware schemes consider criteria from the network, from the end-user, as well as from the device side thus aiming to have the most comprehensive view. Topically somewhat close to the previous survey, Yan et al. [59] classified existing schemes regarding "received signal strength (RSS) based", "bandwidth based", "cost function based", and "combination" approaches. The first class actually represents the well-known, classical approaches for handovers where decisions are based on the received signal strength in both, the originator network, in which the end device originally resides, and the target network. Second, bandwidth based schemes make decisions on criteria trying to model the amount of bandwidth that is available either in the originator network, in potential target network, or in both. Lastly, cost function and combination schemes are comparable to the classification of Kassar et al., whereby the combination schemes include fuzzy logic, neural network-based as well as context-aware approaches.

The argumentation in both surveys clearly identifies that decision schemes relying on signal-strength measures are on the one hand rather simple, while on the other hand important decision criteria may not be covered well. As a result, Kassar et al. favor schemes which consider multiple decision criteria and also emphasize the user perspective. Furthermore, the authors of both surveys argue that schemes based on fuzzy logics or neural networks are rather complex such that it may be hard to include them in an end device with limited computational resources. Yan et al. finally suggest that once the computational performance of end devices will have sufficiently increased, a selected set of decision schemes may be included on each device such that the best one may be chosen in a situation-dependent fashion.

3.5 Recent Mechanisms for Traffic Offloading

The basic categories of offloading traffic from cellular networks to WLANs have been introduced in Sections 3.2 and 3.3, namely immediate and time-shifted approaches. As a basis, standardization bodies of 3GPP have discussed different levels of support from the cellular networks and WLANs regarding their integration.

Existing literature pointed out the state of art regarding different existing integration levels of WLANs and cellular networks allowing to offload traffic by ranging from "unmanaged" up to "integrated" approaches, whereby in the latter case the cellular operator is still capable to offer his services to the mobile end device, while at the same time being able to influence the selection of a wireless access technology [61, 62]. By this, a cellular operator may either unload the cellular radio access network (RAN) or the wired infrastructure of the cellular network denoted as core network (CN), or both. While previous work focussed more on the architectural dependencies of offloading solutions regarding coupling and interworking of WLANs and cellular networks, we jointly summarize the existing approaches from a broader perspective by considering the support from each side towards a seamless offloading, where the mobility management issue

for a handover certainly becomes crucial. In the following, we categorize existing mechanisms and schemes currently under consideration of standardization bodies in categories according to our definition of heterogeneous handovers in Sec. 3.3.1 regarding the responsibility of handling seamless mobility management aspect, namely *mobile-controlled (plus network-assisted)* and *network-controlled (plus mobile-assisted)*.

Mobile-controlled refers to the case that an entity inside the device solely manages the traffic shift from cellular to WLAN access alone—without any support from wireless access providers or other entities. Such shifts of data traffic to WLAN hotspots managed solely by the end device are routed towards the Internet without traversing the CN of cellular operators. In the 3GPP releases for LTE/LTE-Advanced this is referred to as *non-seamless WLAN offloading* [63, Ch. 14, pp. 349–362]. To avoid that the end user has to trigger the setup of an alternative link manually, *connection manager (CM)* applications have been introduced on mobile end devices such as smart phones or tablets for various operating systems (OSs) [64]. For example, the OS Android offers a "ConnectivityManager" [65] that, once WLAN access becomes unavailable, is capable of initiating alternative cellular network access. In the beta stadium, also Apple's iOS 6 offered a functionality denoted as "Wi-Fi Plus Cellular" which was intended to support 3G access if WLAN connectivity is unacceptable [66]. While not included in the final release of iOS 6 [67], however, this functionality was finally incorporated in iOS 7 [68]. Although both approaches come up with the advantage to automatically initiate a setup of an alternative wireless connection, still the current, on-going end to end transport connections, e.g., the Transmission Control Protocol (TCP) connections, break. To resume connections over another wireless access technology as a result of an IP address change, mobile management schemes are required. On IP level, schemes such as mobile IP (MIP) usually rely on a support from network side. Without such an assistance from the network, mobile-controlled handovers have to rely on higher OSI layer solutions which may help re-initiate a transport level connectivity within the alternative wireless access. Note that even some application-level protocols such as the Hypertext Transfer Protocol (HTTP) or the File Transfer Protocol (FTP) may allow to resume a session which however comes with a certain interruption time. Other examples are residing on the transport layer and allow a setup of parallel transport connections over both wireless network technologies thus enabling soft handovers and potentially allowing even a seamless offloading. Representatives for such schemes are the Multipath TCP (MPTCP) [69] or the Stream Control Transmission Protocol (SCTP) [70], for example.

In contrast to pure *mobile-controlled* schemes, also instances in the network may handle or assist mobility management. Such instances in a network can be for example routers in the home network of an end user or gateways in the network of a cellular operator that delivers means for the support of shifting traffic from cellular to other wireless access technologies. Without loss of generality, we stick in the following discussion to the cellular operator example. There, a shift of traffic can be realized either directly by keeping all mobility management functionality inside the CN of the cellular operator or by spreading the functionality among network(s) and the end device side, whereby in the latter case the operator just remains control by specifying and delivering policies to the

end device regarding the traffic steering. Independent of the applied family of mechanisms, 3GPP standardization bodies tried to integrate WLAN hotspots into the cellular network architectures to enable not only access to cellular data services but also to allow a timely shift of traffic between access technologies while remaining full control. For Global System for Mobile Communications (GSM) and/or UMTS, different coupling architectures were standardized, whereby we focus in our short survey on approaches that allow an integration of potential third party WLAN hotspots connected to the CN of the cellular operator via the Internet. The architectural extensions, namely *interworking WLAN (I-WLAN)* [71] and *generic access network (GAN)* [72] are representatives of so-called *loose* and *tight coupling* approaches [73, 74]. The first basically enables the end device to access data services via the network of the cellular operator while being connected with a WLAN cell. This is realized by tunneling the data traffic between the WLAN stack and operator's CN. Besides, I-WLAN still allows to access Internet services directly via the WLAN hotspot, thus being referred to as a loosely integrated scheme. In contrast, GAN requires end devices to use 3GPP-like protocols above IP (and thus above the 802.11 stack), tunneling all traffic into the CN thus essentially integrating the WLAN NIC tightly into the cellular architecture. As a result, a GAN-enabled device may transport not only data but also (circuit-switched) voice traffic over the WLAN link. Ferrus et al. [73] argue that the GAN approach may allow a handover support similar to the UMTS system, whereby interruption times of a handover may be even reduced by utilizing both available NICs concurrently. Further, certain authors pointed out that mobility management approaches such MIP may also enable vertical handovers for I-WLAN architectures [73, 75].

A step further, with the more recent standardization releases for LTE/LTE-Advanced, 3GPP specified the integration of WLAN hotspots into cellular networks for their new all-IP-based CNs. The underlying 3GPP network architecture can integrate *trusted* as well as *untrusted* WLAN hotspots into the CN, whereby the untrusted coupling includes an additional gateway within the CN that is responsible for the handling of tighter AAA aspects [76]. Furthermore, 3GPP standardized two approaches being relevant for our offloading discussion, namely *multi-access packet data network connectivity (MAPCON)* and *IP flow mobility (IFOM)* [63, Ch. 14, pp. 349–362]. The MAPCON approach enables the access of external IP networks (*packet data networks, PDNs*, in 3GPP terminology) via different NICs, whereby the traffic from each NIC is traversing the cellular CN towards the external networks in separate tunnels. Such a tunnel connects a NIC with an external network and is referred to as a *PDN connection*. Each PDN connection however involves a different egress router of the CN, which is essentially the end point of the tunnel and is denoted as *PDN gateway (PDN GW)*. Now, with the parallel access to external networks via separate PDN connections, MAPCON allows to allocate IP traffic flows to PDN connections, effectively enabling a traffic offloading on the granularity of traffic flows, thus however supporting no seamless mobility. As a result, IFOM [77] extends offloading of data traffic to non-3GPP technologies such as WLANs in a seamless fashion for *single traffic-flows*. This is realized by a *single* PDN connection spread over multiple NICs, i.e., separate tunnels terminating at the same PDN GW. With this,

IFOM does not only allow to switch all traffic flows of an end device from one NIC to another, but also enables to use both NICs concurrently—either for *seamless handovers of selected traffic flows* or for *splits of flows* among the available NICs on an end device. For IFOM, 3GPP release 10 borrows a mobile-controlled, network-assisted IP Flow Mobility management scheme from the IETF based on *dual stack MIPv6 (DSMIPv6)*. In their survey, de la Oliva et al. [78] pointed out that although such a MIP approach is mobile-controlled, a cellular operator may specify policies on the device side regarding the allocation of traffic flows to wireless accesses via the *access network discovery and selection function* being described in the next section. In addition to the mobile-controlled, network-assisted DSMIPv6 scheme, 3GPP has discussed also *network-based IFOM (NB-IFOM)* approaches, using network-controlled *Proxy MIP (PMIP)* or *GPRS Tunneling Protocol (GTP)* based mobility management schemes [78, 79].

Finally, in both, the *mobile-controlled* and the *network-controlled* category, handover or offloading processes may not be only assisted or handled by a network of a single operator, but may be further distributed over multiple networks belonging to different ownerships. There, mobility management is handled by the current provider with which the end device is transporting its traffic flows at a given moment. Regarding this class of approaches, relatively few work is available yet. In their recent survey, Zuniga et al. [80] described that the IETF currently discusses and envisions so-called *distributed mobility management* approaches. These are in stark contrast to the classical, centralized approaches of 3GPP cellular networks. Basically, distributed mobility management schemes move away from a centralized mobility management by shifting the mobility handling in a distributed fashion to the access routers of the involved (access) networks, whereby the solutions can be again either controlled from mobile or network side, while each may be assisted from the counterpart. How this broad scope of future mobility management will further evolve, is completely unknown and can be seen as a major research and standardization field of its own.

3.6 Frameworks Towards Resource-aware Offloading

We refer to handovers from WWANs to WLANs aiming at an improved usage of wireless resources as *resource-aware offloading*. The last section focussed on recent extensions from 3GPP and IETF standardization bodies aiming at a support of seamless offloading both from the network integration / interworking as well as from the mobility management perspective. Yet, these aspects do not cover completely all dimensions that could be potentially utilized for a support of offloading policies. In other words, the effort regarding vertical handovers, also including an offloading of traffic to WLAN hotspots, can still be seen to be in its infancy: policies are specified and given by a single operator probably including minor extensions that take some status information of the end device side into account. Such an approach is currently followed by 3GPP standardization with its *access network discovery and selection function (ANDSF)*. Beyond these rather initial approaches, a broader framework has been discussed in the course of IEEE 1900.4 that also covers cooperations regarding possible policy negotiations from multiple operators

of wireless access networks thus essentially moving away from today's paradigm that a single network operator solely decides on his own. This may enable to allocate end devices with their traffic streams to access networks in a resource-optimal way such that the occupied amount of wireless capacity is minimized while still a given QoS level for the end user is obtained. While such potentials for negotiations among operators certainly enable a big step ahead, dedicated support for each technology-specific wireless access link is likely expected to complement the road towards fine-grained, policy-based decisions. Such a support enables a fast information of upper OSI layers about link changes or timely triggers of the handover process. This aspect is covered by the IEEE 802.21 initiative with its media independent handover framework which however requires a proper interfacing down in each wireless access technology. In the following, we shortly describe the relevant aspects for traffic offloading from the ANDSF, IEEE 1900.4, as well as IEEE 802.21 perspective, whereby we note that the two latter were combined conceptually by Dimitriou et al. [81] to a joint framework already.

3.6.1 3GPP ANDSF Framework

3GPP standardization came up with the ANDSF [82] as a means for cellular operators to specify policies regarding a control of non-3GPP access for the end devices of their customers. These policies have been grouped by 3GPP into three categories which are namely *access network discovery information (ANDI)*, *inter-system mobility policy (ISMP)*, and *inter-system routing policy (ISRP)*. ANDI delivers information about the available heterogeneous neighbor networks to end devices on their request. This information includes a description of each technology being within the range of the end device, a list of available cells for each, and more technology-specific details, e.g., frequency bands or certain configurations. The second, ISMP, is intended for end devices that are either not capable or not configured to transport IP traffic simultaneously over multiple NICs. The ISMP specifies not only whether a change of the radio technology is allowed from the operator perspective, but further selects the most suitable wireless technology and the related wireless cells to access the CN of the operator. Finally, ISRP governs the set of rules for the cases in which an end-device may simultaneously transport IP traffic over multiple NICs, i.e., for MAPCON, IFOM, and non-seamless offloading as introduced in the previous section.

From the architectural perspective, the ANDSF is a *network element* inside the CN of a cellular operator [82]. It delivers the corresponding policies to a connection manager entity on the end device that acts upon the intended behavior. While ANDI can be just requested by the end device, ISMP as well as ISRP may be either requested by the end device (*pull mode*) or sent by ANDSF to the device (*push mode*), e.g., as a reaction to network events. The frequency for ANDSF updates has not been specified by 3GPP, nevertheless a minimum time interval is foreseen between two updates thus allowing to cap the resulting overhead.

3.6.2 IEEE 1900.4: Towards a Distributed Decision Making

The IEEE 1900.4 standard is a high-level architectural framework defining "Architectural building blocks enabling network-device distributed decision making for optimized radio resource usage in heterogeneous wireless access networks" [83]. Coming from its original background of cognitive networks, it considers fields of application regarding the delegation or the sharing of spectrum among networks or end devices. In addition to these aspects, it further targets a "distributed radio resource usage optimization". As for example pointed out by Dimitriou et al., this standard is certainly relevant also for vertical handovers [81], and may apply also to network-centric, resource-aware schemes. With its generic components, IEEE 1900.4 is intended for scenarios including decision making processes also among *multiple operators*, thus being far ahead of the actual technology-dependent standardization process of 3GPP or IEEE 802.11. IEEE 1900.4 introduces architectural entities on the operator as well as on the end-device side. The most important entities are thereby so-called *reconfiguration managers* on both sides. The "network reconfiguration manager (NRM)" is responsible for managing and deriving optimization criteria regarding the resource usage of the whole heterogeneous wireless network including the devices. As such, the NRM may be also distributed among several network entities. In contrast, the "terminal reconfiguration manager (TRM)" on the end device is responsible for an optimization of the local conditions on the device or link level, thereby following the constraints of NRM, user requirements and preferences [84]. Both, NRM and TRM interact via a "logical communication channel" that is denoted as "radio enabler (RE)" [83]. Managers on both sides are assisted by measurement and controller functionalities, which provide any relevant measurements for the decision processes and conduct the actual (re-)configurations on each side. The reader is referred to the survey of Buljore et al. [84] for an in-depth description of these entities.

We shall note that a single-operator scenario in IEEE 1900.4 may come quite close to the current ANDSF standardization works of 3GPP, where the network-based ANDSF specifies policies that are translated and executed by an entity on the end device. However, the proposed architectural framework by IEEE 1900.4 explicitly includes also multi-operator scenarios. For this, two fields of application are given [83]. In the first, one NRM exists per operator with an interface for inter-NRM communications thus enabling a direct interaction among each other. In contrast, for the second, all operators jointly exhibit one external NRM which orchestrates all involved heterogeneous networks. Yet, to enable a realization of these envisioned fields of application, the standardization bodies are working towards precise descriptions of interfaces and protocols for the high-level IEEE 1900.4 architectures within 1900.4.1 [85, 86].

3.6.3 Media Independent Handover Framework: IEEE 802.21

For more than a decade, one of the superior challenges for vertical handovers has been to avoid a degrading QoS for traffic flows to be shifted from one access link to another. This QoS issue appears both as a result of interruption times as well as packet losses due to switching the link as well as well as changing the IP path [88]. The latter is usually

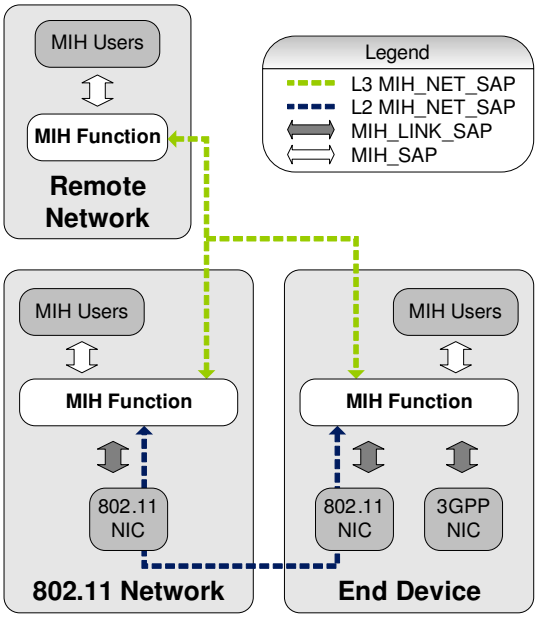


Figure 3.4: Overview about MIHF interworking, modified after [87, Fig. 5, p. 26]

handled by mobility management schemes either on IP level or above. Thereby, isolated approaches on single OSI layers have shown to be not able to handle the corresponding issues on a satisfactory level. The general problem lies in the fact that a vertical handover involves different layers of the OSI model ranging from the link layers, up to the network or even higher layers. By design, the OSI model introduced the layer concept such that a lower layer hides the detailed complexity of its task by presenting only an abstraction to its higher layer, essentially by offering a specific service over well-defined interfaces [7]. However, each specific step for a handover in each layer requires a certain amount of time, e.g., neighbor network discovery and link setup at layer 2 or the re-establishment of an IP path at the network layer, and may further depend on each other in the time domain. Thus, towards handovers it is highly beneficial to handle the corresponding steps in a more orchestrated fashion among different layers [88]. For example, one may want to be capable of informing higher layers quickly about link level changes or prefer to control lower layer behaviors tightly on a higher level, effectively enabling make-before break handovers with timely preparations, if possible in parallel on the different layers.

The IEEE 802.21 standard [87] tackled the above issues with the design of a generic architectural framework aiming to assist network selection and handover decisions in heterogeneous networks, whereby the specification of policies as well as decision entities remained out of scope. The main component of this architecture is the *media independent handover function (MIHF)* that conceptually resides inside a protocol stack above the link layer. For multi-standard devices with NICs of different wireless technologies the MIHF has an interface to each. Furthermore, a MIHF entity may be placed both on end

devices and on the network side(s), e.g., on a WLAN AP or on a router, as shown in Fig. 3.4. IEEE 802.21 has foreseen services not only of a MIHF inside a protocol stack of a single device, but further offers possibilities for a *remote* handling of MIHFs on different network devices. For both, local and remote applications, the MIHF offers three services to the upper layers, denoted by the standard as "media independent handover (MIH) users". These MIH users are usually the protocols on the upper layers being responsible, e.g., for the mobility management or the execution of policies. To these MIH users, the MIHF offers *the event, the command, and the information service*.

Let us briefly discuss each of these services defined in [87]. Events are basically triggers indicating either an occurred or an expected change in the wireless link conditions at the lower layers, i.e., PHY and MAC of the concerned wireless technology. The MIHF passes these pieces of information to the MIHF user(s), thus offering the *event service*. In the remote case, the "local MIHF" forwards the event to its "peer MIHF" that further gives it up to the "remote MIH user". In contrast, the *command service* gives the MIH users the possibility to control locally the behavior of lower layers and remotely to steer higher as well lower layers. These commands comprise to trigger handovers, to start the neighbor discovery and link establishment process with a selected layer 2 network, and others. For this service, the MIHF on the executing side forwards the commands to the responsible layers. With its local and remote functionality, this service is capable to support network-initiated as well as mobile-initiated handovers. Lastly, *the information service* enables the MIH user to obtain various pieces of information about available wireless neighbor networks. This includes the distribution of rather permanent settings such as the basic level of QoS support, the security requirements, or the options regarding a support of mobility management schemes by this network. Further, information may be obtained regarding the actual configuration of the network—e.g., with respect to the frequency band in use, thus minimizing the effort for the discovery of networks. Finally, we shall note that the delivery of dynamic aspects from neighbor networks by the information service is out of scope of IEEE 802.21 standard and is left to the end device.

One important feature of IEEE 802.21 is the flexibility regarding the remote application of the services. This remoteness is realized on two levels via different *service access points (SAPs)* as highlighted in Figure 3.4. First, the L2 "MIH_NET_SAP" allows the interoperability of MIHFs via the wireless access technology, in the given example via the WLAN link. Second, MIHFs may be placed elsewhere on other network components not equipped with wireless NICs. These components, such as servers making handover decisions or offering neighbor network information, may be placed anywhere in a wired access or a core network, e.g., also in the CN of a cellular operator. Following the 802.21 notation, we refer to these networks as *remote* networks or network parts, since they do not have any access to the wireless technology inside their local protocol stack. Interworking between "remote MIHFs" and local instances are realized on IP level via the L3 "MIH_NET_SAP". From the IEEE 802.21 perspective, it is this connectivity between MIHF entities on the wireless APs as well as on the remote network components that may allow for a collaboration and an interaction of different operators regarding joint network selection and handover decisions.

Finally, we shall note that the IEEE 802.21 concept from its early beginning has gained considerable attention in the research community regarding diverse application areas. Ghahfarokhi and Movahhedinia [89] summarized the different areas in which 802.21 has been mainly applied. The survey points out that existing work mainly focussed on improvements of mobility management schemes by 802.21, events regarding flattering wireless links, handover decision schemes, and extensions to the information service enabling a support of more dynamic aspects about available neighbor networks. The work by Silva et al. [90] further discussed the integration of the 802.21 MIHF into the Android OS for end devices such as smart phones. As an MIH user on the device side, the authors introduced an "Android Mobility Manager" on top of the MIHF. Among others, this manager controls the mobility handling at layer 3 by guiding a local MIP instance. The authors evaluated their solution by measuring packet losses and interruption times for WLAN/3G handovers in a testbed for various traffic types. Their results indicate not only a strong improvement in combination with 802.21 over legacy MIP handovers, but also identify that for time-bounded services such as VoIP, although the QoS degrades, the perceived quality by the end user may stay on an acceptable level—with interruption times of around 150 ms and packet losses of up to five percent for VoIP traffic.

3.7 WLAN Radio Resource Measurement and Network Management: Obtaining Technological Parameters

In its early years, the 802.11 standard did not include a detailed support regarding the management of a WLAN network and its radio resources. Although the associated number of end devices could be controlled by the network side, an important piece of the puzzle for fine-grained decisions was lacking. Decision entities, either residing on a WLAN AP or even deeper inside the wired part of the access network, had no possibility to gain detailed measures about the state of the WLAN environment—from AP level as well as from the end device side. As such, two amendments to the 802.11-2007 standard [91] were developed, namely 802.11k and 802.11v[2], tackling the above issues as explained in the following.

3.7.1 IEEE 802.11k: Radio Resource Measurement

The 802.11k amendment "Radio Resource Measurement of Wireless LANs" [92] filled a significant part of this gap by introducing a set of measurement procedures consisting of standardized metrics, schemes for controlling the measurements, and procedures for the signaling of the measurements as well as other information. Local measurements at a given piece of WLAN network equipment consisting of an 802.11 PHY and MAC may be remotely requested by others, both from end devices as well as from a WLAN AP. For the remote handling of measurements as well as the signaling of other information, 802.11k uses *request and report handshakes*. For measurements, the requests include the metric(s)

[2]802.11k and 802.11v were an amendment to the 802.11-2007 standard. Both are now inherently included in the most recent version, 802.11-2012 [6].

as well as the start time and the selected duration. By this, the peer device is triggered to measure for a pre-selected amount of time and to send a report afterwards. Further, a report has to be generated on an event basis, if a predefined threshold of a certain metric has been crossed. In addition, 802.11k also extends the list of so-called 802.11 "information elements" which are essentially small modular containers including a dedicated piece of information. These information elements may be attached to the 802.11 beacon frames thus informing also non-associated devices about the conditions in the WLAN cell.

Both, 802.11k as well as 802.11v detailed below, were also designed to fit into the 802.11 device management plane introduced in Sec. 2.2. For this, the MLME MIB (compare Sec. 2.2) was extended by additional data structures for requesting and storing radio resource measurement and network management information. Following the notation of 802.11k, we refer to this specific part of the MIB as the *radio resource management (RRM) MIB*. This RRM MIB contains not only local information of an 802.11 device, but it also capable to store obtained data from remote STAs. Finally, via the 802.11 SME interface, MAC measurements and management information may be requested by higher layer entities.

In the following, we shortly give an overview about selected features of 802.11k being relevant for this thesis. The interested reader is referred to the original 802.11k amendment text [92, pp. 111–132] for the exhaustive list with its in-depth description. For a brief survey, we categorize selected 802.11k means regarding the applicable range of this thesis, namely the neighbor network discovery, the transmission statistics of a selected device, and the load level of a WLAN cell. For the neighbor network discovery, 802.11k offers two distinct means:

- The "Beacon Report" feature enables to trigger a selected WLAN device to conduct a search for surrounding WLAN cells and to obtain information about the received signal strength levels for each. Several flavors of these reports enable a broad range of applications. They may be used to find pre-selected WLANs identified by their *BSS identifier (BSSID)*, to search for WLANs just on a subset or on all available channels, and to signal already existing information about neighbor cells without actually triggering measurements.

- The "Neighbor Report" is requested by WLAN end devices from an AP. Dependent on the request, the report contains either a complete list of all known neighbor WLAN APs or a selected list that includes APs being members of the same or a different ESS. Further, the report provides information about each AP regarding its configuration and security settings.

- Lastly, an AP may assist the network discovery process by including an "AP Channel Report Element" in its beacons. This element essentially indicates on which channels other WLAN APs might be found.

Gathering statistics regarding the transmissions of a selected WLAN device is tackled by the following 802.11k options:

- The "STA Statistics Report" is reflecting the view of a WLAN device regarding its transmitted and received frames. This report includes various measurements ranging from the counts of successful and failed transmission attempts, the number of conducted retransmissions to the number of received frame duplicates and missing ACKs. On AP side, it additionally comprises the number of associated devices, the load level at AP, and the mean access delay for all downlink transmissions and the transmissions per "traffic category".

- Further enabling statistics not only per device but also per traffic category or even per traffic flow, the "Transmit Stream/Category Measurement Report" extends the STA statistics on a per-flow / per-traffic-category level. With the request/response scheme describe above, a device is triggered to measure for a pre-selected amount of time and to send a report afterwards. These statistics may also be configured for a "triggered autonomous reporting". In this case, the STA generates on its own a report if a predefined event happens, e.g., a threshold of a given metric has been crossed.

Information about the actual load situation in a WLAN cell was already included before 802.11k. The 802.11-2007 standard [91] already described the "BSS Load Element" reflecting not only the number of associated devices with an AP, but also giving the amount of time that the AP has perceived a busy channel, determined by the 802.11 carrier sense. 802.11k further describes additional measures giving more detailed insights about the load situation:

- The "Channel Load Report" allows to obtain load information from any WLAN device. Thereby, load is determined as the fraction of time regarding a given measurement interval that the measuring device has perceived the wireless channel to be in the busy state.

- The "BSS Average Access Delay Element" indicates the amount of time that an AP requires to conduct its downlink transmissions. This time span starts when the MAC takes the data frame out of the queue and ends when it receives a corresponding ACK frame. If the AP is conducting medium access according to the 802.11e EDCA, then the access delay is averaged over transmissions of all the traffic categories. Further, for the mean access delay per traffic category, 802.11k introduced the "BSS AC Access Delay".

3.7.2 IEEE 802.11v: Network Management

The 802.11v amendment "Wireless Network Management" [93] on the one hand extends the procedures of 802.11k, but also incorporates schemes for an improved management of homogeneous networks, e.g., with the purpose to balance the load among WLAN cells belonging to one ESS. As such, 802.11v focusses mainly on a homogeneous network

management. On the other hand, it opts for a certain support allowing to obtain information about other available, heterogeneous NICs of an (end) device. In the following, we discuss the pieces of the 802.11v amendment that are highly relevant for this thesis, whereby the interested reader is again referred to the original standard amendment for an exhaustive description of all options an extensions regarding the WLAN network management [93, Chapter 11, pp. 229–307]. Regarding 802.11v, we focus on three aspects, namely improving transmission statistics of a selected device, obtaining "vendor specific information" about the WLAN NIC in use, and getting information also about heterogeneous access capabilities of a device.

Starting with the transmission statistics of a selected device, the 802.11v domain basically extends 802.11k's "STA Statistics" by enabling the "triggered autonomous reporting", i.e., a "STA Statistics Report" is generated if the threshold of a given metric has been crossed. In standard terminology this is denoted as "Triggered STA Statistics". This feature enables a report on an event basis, thus circumventing to specifically trigger "STA Statistics" multiple times and in turn reducing the signaling overhead.

For the second and the third aspect, obtaining information about vendor specific issues as well as heterogeneous capabilities of the WLAN device, we utilize the "Diagnostic Reporting" procedure of 802.11v. For this type of request / report handshakes, the amendment introduced the "Manufacturer Information STA Report". Among others, it includes a manufacturer identification, the NIC model, its firmware version, as well as an information about the actual device type (separating between notebooks, multi-standard smart phones, and others) and other available radio access technologies on this device.

Chapter 4

Challenges and Scope of Thesis

Offloading traffic from a WWAN access to other, small-coverage cells such as WLAN hotspots is seen as the de facto solution both from a technical as well as an economical point of view to deal with the traffic increase that is expected to roll-over WWAN operators in the next years. As pointed out in the previous chapter, standardization efforts currently take place to enhance existing offloading approaches aiming at a support towards a *seamless* migration of end users' traffic. The term 'seamless' thereby refers to handovers of running traffic streams in a way that the end user does not perceive a degradation in the quality of the specific service. Such types of offloading enables handover decisions at all times to smoothly react to changes in the given conditions, e.g., spikes in the offered load to be served by a network, mobility of (offloaded) end-users, or changes in the network structure due to switching off equipment for energetic reasons. Besides the efforts that we discussed in the preceding background chapter, nevertheless a broad field of challenges remains for a support of such offloading. Especially it is challenging to maximize the benefit of using small cell WLANs operated in enterprise, hotspot, or end-users' home scenarios. In a nutshell, what has not been included in offloading decisions so far are aspects regarding a resource-aware offloading considering the target WLAN. This chapter firstly discusses and categorizes remaining challenges in the context of such offloading, presents our reference scenario together with basic assumptions, introduces our underlying architectural framework, and argues about the selected scope of this thesis.

4.1 Limitations of WLAN Network Selection

Still with all the existing or envisioned approaches, offloading decisions are not based on accurate, systematic insights of the WLAN target network. Once powered up, WLAN devices are usually faced with a density of WLAN hotspots such that they can select one out of several APs for an association, which is known as *AP selection*. As far as the 3GPP standardization is concerned, a *connection manager (CM)* on the end device is expected to make the network selection decisions being provided on the one hand with the policies

of the WWAN operator and on the other hand with some information about neighbor WLANs. For the latter type of networks, a device obtains knowledge about capabilities and initial RSSI measures during the neighbor network discovery. Overall, it is an issue how network selection in 802.11 networks happens today. Sophisticated approaches can be found rather in enterprise WLANs. There, a group of WLAN APs belonging to the same administrative instance is usually orchestrated by a central AP coordinator. This *AP controller (APC)* may also decide for the association of devices with a specific AP as discussed by Murty et al. [94]. Contrary for smaller WLAN deployments, this network selection process is usually terminal-centric, i.e., the WLAN end-user device selects the WLAN AP to which it will connect to, and bases on very simple measures such as the RSSI level. Note that this selection is not specified by the 802.11 standard leaving a broad space for proprietary, vendor specific solutions. Thus, whether a network selection scheme is actually good or bad in terms of an efficient usage of WLAN resources broadly depends on the proprietary scheme applied on each single WLAN device.

However, one could think about a support of the network selection decision on device side by delivering more and more detailed information from the WLAN hotspot. 802.11k, for example, enables a signaling of the current cell load perceived by an AP, thereby providing the ground for more sophisticated decisions on STA side [8, 95]. By this, Abusubaih et al. [8] associate devices with WLANs following the goal to distribute the load over several cells in a smooth manner trading off achievable device throughput and the impact on other present WLAN STAs. With a similar approach for centralized admission control decisions on the AP, Rossi et al. [96] consider predictions regarding the obtainable throughput of a device that has joined a BSS. Additionally, they take into account how an added STA may affect the whole cell as well as other, already associated STAs in terms of throughput. Increasing the scope of such information may help further, however we expect the improvement of such schemes to be limited as also the *knowledge* of an AP about its surrounding environment is highly limited, too, which in turn affects also such AP-centric approaches in enterprise WLAN networks.

The major issue with 802.11 networks essentially lies in the fact that they operate in *non-exclusive* 2.4 and 5 GHz ISM bands. Today, the most common 802.11b/g/n WLANs are actually applied at 2.4 GHz, where at most three non-overlapping WLAN channels are possible in the frequency space at all. However, the increasing popularity of the WLAN technology over the last decade has lead to a density of WLAN APs exceeding the available non-overlapping channels by magnitudes. As a result, a WLAN hotspot today has potentially to deal with a large number of interfering WLAN cells on the same or adjacent frequency bands. Even worse, the crowded 2.4 GHz band is used by many different administrative domains, i.e., different owners of hotspots, such that 802.11 technology is faced with lots of non-controllable interference. Note that 802.11 follows the strategy of a local radio resource management, which is conducted for managing the own domain. A joint management or a collaboration with other administrative domains has been out of scope of the 802.11 standardization so far. Thus it remains questionable to what extend future network selection schemes for both, small WLAN deployments as well as enterprise networks, will incorporate sufficient accurate and timely knowledge

of all these aspects really enabling optimal decisions. Further, while multi-cell WLANs from one administrative site may reorganize associations of devices to APs, still a certain subset of these devices may suffer from suboptimal selections or associations because they lack alternative homogeneous access options.

4.2 The Way towards Resource-aware Onloading

Offloading results in a handover directed from the WWAN to a third party network such as a WLAN hotspot. However, enabling one-way decisions about traffic streams to be offloaded to WLAN falls short if one aims at a maximization of offloaded traffic. From the context of today's discussions about offloading, it seems that such a maximization is expected to play an important role towards a solution for the high traffic demands. In contrast to this perspective, envisioned network selection mechanisms as provided by 3GPP's ANDSF employ coarse information about candidate networks, i.e., statements about the basic availability of WLANs on the basis of neighbor network information. Recent proposals discuss more flexible enhancements to the ANDSF providing a neighbor network discovery based on feedback from the end-user device and a network selection that considers different user patterns and their requirements [97]. Although this enables the possibility of dynamic decisions to some degree, it still falls short regarding a true resource-aware allocation of devices to wireless accesses. The basic problem lies in the following issue: when a WLAN network selection decision is imminent, it is usually not known how a device with its traffic stream(s) will behave in the selected target network. As discussed in the preceding section, the initial network selection processes usually base on some simple heuristic which in turn may lead to good or bad decisions. From the offloading perspective, bad decisions are the ones which are costly both in terms of wasted resources as well as a degraded quality perceived by the end user. An unnecessary amount of resources is not only spent in the network to which the traffic is offloaded to, but also in the WWAN of an operator. The reason for this is quite simple: suboptimal decisions occupy resources in the WLAN cells. Such resources could be actually better used by other traffic flows being also candidates for offloading decisions. We conclude that a key piece of the whole offloading puzzle has still not been discovered yet: especially in loaded WLAN scenarios, it is indispensable to identify suboptimal network selection decisions that actually waste resources in WLAN as it turn may also improve the overall situation in WWANs.

What is actually needed is a kind of fallback option once 802.11 access is not available or not suitable anymore. As the mobile WWAN operators are known to support not only high mobility but also support delay- and loss-sensitive applications such as voice calls, the WWAN is expected to take over such traffic again if the end user becomes mobile and leaves the cell. We denote such a fallback option as *onloading* throughout this thesis.[1]

[1]For the sake of completeness, we note that the term of 'onloading' traffic to WWANs has been recently used in a different context in the literature in [98,99]. In contrast to this work, these approaches temporarily expand the available capacity for users in a residential digital subscriber line (DSL) network by *additionally* using 3G/4G cellular accesses of multi-standard end devices thus enabling a multi-path transport for selected traffic.

As discussed in Sec. 3.5, today's end-user devices already incorporate fallback options such as iOS's "WiFi Plus Cellular" that enables cellular access if WLAN connectivity is either not capable to enable access to the Internet or does not deliver a desired QoS level. However, such device-centric fallback strategies fall short in the context of our discussion. Similar to such a fallback option, i.e., supporting a handover of offloaded traffic back to WWAN cells, we advocate such *onloading handovers* also for other end-user devices in WLAN. Specifically, we propose an onloading of devices which have become noticeable regarding a rapid waste of WLAN resources being a result, e.g., of interference problems with other WLAN cells, of device behavior such as sub-optimal link data rate adaptation, and others.

As we expect customers of WWAN operators basically to pay in the future rather for the delivery of content together with a continuous service at a given quality than being charged in terms of called minutes or transported amount of data, following this rationale will confront such operators with the need to help offloaded end-user devices also to be onloaded again, if possible even in a seamless fashion.

4.3 Challenges from the Hotspot Perspective

As discussed above, suboptimal network selection decisions for traffic offloading may impose problems for the end user, for the selected WLAN hotspot, and for the mobile operator. Thus, what we called 'fallback option' and 'onloading' above is essentially nothing else than a handover back from WLAN to the WWAN part. Following such a rationale, a couple of challenges emerge in different dimensions. For an overview, we briefly discuss them.

4.3.1 Cooperation Between Owners of Administrative Domains

The emerging trend for heterogeneous handovers across networks owned by different operators imposes great challenges as these ownerships may require a *direct* interaction and coordination between the owners of the involved administrative domains, especially when aiming at a resource optimal allocation of traffic streams to networks. In today's situation, WLAN hotspots can be operated at many different levels ranging from private owners applying WLAN at home or in their small businesses such as cafes or shops, over enterprise WLANs covering larger areas such as whole companies or campuses, up to WLANs being operated by WWAN providers and thus being integrated tightly into their networks. One can expect a large space for improvements if the allocation of traffic flows to different wireless access cells is not only based on the local, rather limited view of either the end device or a single access cell. Further, such decisions may incorporate dynamic knowledge from all the network entities managing the involved access networks. For example, this involvement shall not only include the candidate cell for a handover but may also consider other hotspots or even other (heterogeneous) technologies being operated in the direct vicinity and using interfering frequency bands. Obviously such a global view, at least in a given vicinity, is very challenging while aiming at the same

time for a scalable and efficient operation both in terms of the number of end devices as well as the number of involved networks. Visionary concepts such as the "Connectivity Brokerage" framework [100], among others, actually aim at such a well-coordinated network selection nevertheless also advocating a "structured and formalized approach" to keep complexity of a complete cooperative system on a controllable level. Essentially, the support of such an approach will still remain the major challenge especially for the future standardization efforts as current heterogeneous types of technologies, stemming from different standardization bodies such IEEE and 3GPP, do not seem to fully tackle equitable interworking towards such a cooperation yet. Thus, besides all the efforts, e.g., in the IEEE 802.21, IEEE 1900.4, and 3GPP ANDSF work as described in Sec. 3.6, it remains questionable whether end devices as well as network equipment will support such cooperation approaches at a near point in time.

4.3.2 Mobility Management Schemes

For offloading traffic from WWANs, one actually has three fundamental options: to either unload the radio access, the wired core network of the WWAN technology, or both. With its IP flow mobility approaches aiming to enable seamless traffic offloading, the current 3GPP standardization for cellular networks just tackles the first option, while still keeping all the related functionalities regarding a mobility management inside the core network of the cellular operator. Note that this type of mobility management imposes several drawbacks, ranging from routing to scalability issues, as data packets are usually treated by a centralized management entity, which in turn has to handle the traffic of a large number of end users while being a "single point of failure" [80]. Even worse, if one aims to reduce the load in the core network by traffic offloading, there is no standardized solution today for next generation 3GPP cellular networks to support this in the context of seamless handovers as the centralized mobility management of the core network is obviously not available anymore. However, for a smooth offloading to WLANs on a per flow level, the feature of seamless mobility would be badly needed. How to support this while getting most out of offloading approaches, especially for the core of a cellular network, has yet not been standardized by 3GPP. Light at the end of the tunnel seems to appear from the IETF world in the context of *distributed mobility management* schemes [80], which have been discussed in Sec. 3.5. Again, how this broad scope of future mobility management will further evolve, is completely unknown and can be seen as a separate major research and standardization field, to which this thesis will remain orthogonal. Throughout this work, we just assume the presence of a mobility management scheme handling the switching of traffic streams between the WLAN and the WWAN.

4.3.3 Support of Single-Radio Handovers

Over the past years, the design of end-user devices such as smart phones or tablets has been driven on the one hand by a miniaturization for a flexible handling and on the other hand by energy efficiency considerations to endure the lifetime for a battery-driven operation. Both aspects have been also highly influencing the communication capabilities

of such mobile devices leading to NICs with chips and transceiver chains supporting a group of different wireless access technologies ranging from cellular networks to WLAN, Bluetooth, and sometimes even WiMAX. We denote a single radio interface supporting multiple network technologies as a *multi-standard NIC*. Such an "integrated radio design" imposes constraints in two different dimensions [101]: First, heterogeneous technologies on a multi-standard NIC may impact each other due to the radio designs, where the actual hardware parts are placed very close to each other. Due to this, transmissions in the first may affect a link in the second technology on such a NIC, even impacting non-overlapping frequency bands. Second, putting an integrated design of such multi-standard NICs to an extreme finally leads to a shared usage of certain hardware parts by different wireless technologies, e.g, of the transceiver chains, such that only a single link can be active at all a given instance in time. In this case, we refer to them as *multi-mode NICs*. These limitations impose severe drawbacks for a support of handovers at layer 2, especially for delay-constrained services. The preparation of such handovers includes several steps regarding the neighbor network discovery, the decision about the availability of a certain technology for a given type of end-user traffic, the selection of a proper access point, and finally the completion of the association procedure. All together, this can take a significant amount of time. Thus, it is hardly recommendable to enforce a hard handover, i.e., breaking the ongoing communication association before assuring that another, better connectivity is really available. Realizing a soft handover with service continuity is restricted, however, by the applied communication hardware. A support of soft onloading handovers from a WLAN hotspot either to other WLAN cells or even to non-cellular heterogeneous wireless technologies on highly integrated multi-mode NICs would allow to distribute selected traffic among available access cells. Again, note that only a single wireless connection can be active over such multi-mode NICs. How to support a soft handover under these circumstances is denoted as *the problem of single-radio handovers*. Enabling single-radio handovers not only for homogeneous WLANs but also for heterogeneous technologies sharing hardware components on the same NIC has yet been a challenge up to recent years, not only for elastic but also for a support of delay-constrained traffic. As the miniaturization and the energy constraints of end-user devices are likely to be pushed to newer limits, we identify the preparation of single-radio handovers as one of the major challenges in the context of the onloading discussion for next generation wireless networks aiming for a proper shift of the traffic load.

4.3.4 Handover Candidate Selection in WLANs

Loading up WLAN hotspots with a maximum number of devices on the one hand, while taking care on the other hand that the QoS of each traffic stream still upholds a certain level, first requires so-called admission-control schemes. A diversity of different admission control schemes [102,103] for 802.11 networks have been proposed over the last decade aiming to avoid an overload of the cell by granting wireless access to too many traffic flows. In this context, the 802.11e amendment from 2005 was a key step enabling a differentiated handling of traffic classes in WLAN hotspots. In their survey, Gao et al. [102] distinct between "measurement-based" and "model-based admission control",

depending on how metrics for admission control decisions are obtained. Among others, the summarized approaches apply various metrics reflecting the occupied versus the remaining, free time of the wireless channel, collision levels, QoS-dependent contention parameters and the throughput of each device. More recent work further considers metrics regarding the delay of packets either as a result of the medium access [104] or in combination with queuing effects as well [105].

While admission control schemes prevent a wireless cell from being overloaded, still, in the context of the offloading discussion, an important issue remains. When aiming at a maximization of offloaded traffic, a resource optimal allocation of traffic flows to the WLAN hotspot remains challenging even in the presence of such admission control schemes and their metrics. Tuning the amount of offloaded traffic from WWANs additionally requires the identification of devices or even traffic flows that are not resource optimal, thus wasting wireless resources and in turn degrading the performance of the whole cell. Back in 2003, Heusse et al. [106] already showed that a single WLAN device transmitting at low link data rates degrades the throughput of all other associated devices. Besides the link data rate selection, the retransmission level as a result of collisions and interference as well as flattering channels have a strong impact on the resources that a WLAN device occupies in a hotspot.

In this context, heterogeneous handovers from WLANs back to the WWAN aiming at a refinement of occupied resources may develop as one of the dominant means to effectively deal with the ever increasing amount of traffic that is forecasted to be offloaded to WLANs. As discussed in Sec. 3.4.3, simple decision criteria consider single WLAN parameters such as measured RSSI values. Yet, a measure has been lacking to combine all these link-layer aspects to a performance metric adequately reflecting the utilization of occupied resources by single WLAN devices, thus allowing to select candidates for an onloading, i.e., a handover back to WWANs.

4.3.5 Estimating the Behavior of Traffic Streams to Be Offloaded

Once an end-user device has been associated and authenticated, WLAN hotspots usually enable wireless access without having a detailed knowledge about the traffic streams to be transmitted. Following the concept of a maximization for traffic to be offloaded to WLAN cells, it would be advantageous from the WLAN perspective to have more insights not only about the end-user applications and their resulting traffic flows prior an offloading, but also about the behavior of the WLAN protocol stack on the end device regarding vendor-specific mechanisms. Such knowledge on WLAN AP-side may rapidly improve admission control algorithms, which regulate the access to the hotspot by avoiding overload situations.

Being able to map traffic streams to distinct applications on the network side may further enable an understanding both about occupied resources and the QoS requirements of a stream. However, some traffic streams are difficult to categorize, e.g., traffic of HTTP applications [107]. The classification of traffic has been an active research field over the past decade with recent advances also allowing for an estimation of the characteristics of each traffic stream such as the overall amount of transmitted data or its duration [108].

Nevertheless, even if an AP would have dedicated knowledge about an up-coming traffic stream regarding the behavior in terms of its duration or its data volume, still its influence on the whole layer 2 specific mechanisms is yet an unknown issue. Even for a traffic stream that has been recently shifted to a WLAN cell, the collision and the retransmission level is completely unknown on longer time scales up to several seconds. While network selection schemes usually try to tackle these issues analytically on the basis of the well-known Bianchi model [109] and its extensions, such approaches fall short in practical wireless networks, where the behavior of the end-user devices do not only depend on the standardized 802.11 mechanisms but also adhere to vendor-specific implementations, parameterizations or policies. How to enable a forecast about the behavior of a traffic stream being selected for an offloading to a WLAN cell even in this practical context has remained an unexplored issue yet.

4.4 Description of the Reference Scenario

In this section, we present our reference scenario together with related assumptions. Afterwards we review the selected assumptions by discussing related work. Then, the following sections survey our underlying architecture and finally argue in this context about the selected scope of this thesis.

4.4.1 Scenario with Baseline Assumptions

From the networks' side, the scenario consists of a set of N WLAN hotspots and a WWAN using a complementary technology in terms of coverage, mobility support and QoS for the end user. We assume that the WWAN covers an area which includes all N 802.11 hotspots. Regarding the selected technology for the WWAN, we consider two options throughout this work. The cell consists either of 3GPP cellular technology such as 3G/4G or is composed of IEEE 802.16e WiMAX equipment. We further assume for both a wired network backplane with a connectivity to the Internet. The exact technology selection is given later at the beginning of each chapter. Finally, we assume that the operator of the WWAN enables to transport as well as to off- and onload all packet based services, also including traffic of delay-sensitive services such as Video or VoIP applications.

Regarding the WLAN hotspots, we assume that they are operated and managed independently from each other, while the coverage area of at least two WLAN hotspots overlaps in space. Each WLAN hotspot has a wired connection to the Internet, whereby we assume that the wireless access part is the bottleneck in our reference scenario. Finally, 802.11 hotspots operate according to the basic PHY and MAC functions described in Chapter 2. Information about the configuration of the hotspots is given later in each chapter. For a hotspot in our setting, we assume an underlying architectural framework that is detailed below in the subsequent section.

Throughout this work, we assume that M end users have wireless connectivity to the Internet by means of multi-standard devices such as a smart phone or a tablet as well as larger pieces of equipment such as laptops. These multi-standard devices support 802.11

as well as the selected technology of the WWAN and are capable to authenticate with all considered networks. In addition, we assume that end users do not move during a communication session, i.e., they have a nomadic mobility. Further, we presume that a significant fraction of end users' traffic flows is long-lived in the sense that they last at least a couple to some tens of seconds.

4.4.2 Discussion of Assumptions

Movement pattern of end users with active communication sessions While being on the move, end users may have active communication sessions specifically by means of hand-held devices such as smart phones or tablets. As such, a mobility of end users certainly plays an important role in the context of handovers. However, we do not focus on this aspect, as we assume that these cases are handled in a smooth way for the end users by WWANs as they have been predominantly designed, from their historical perspective, also for traveling users even at high speeds such as vehicular mobility. In contrast to these mobile cases, the study of residential places from Maier et al. [110] showed that a fraction of end users applies "hand-held" devices solely or in addition to their desktop computers and laptops at home. In other words, these users do not rely completely on their permanently installed options. During an 11-month measurement campaign back in 2008 and 2009, shortly after the smart phone wave emerged, the authors showed an almost two-fold increase regarding hand-held devices accessing the Internet via residential WLAN/DSL accesses. Being able to observe 20,000 DSL accesses, the authors found about 3 percent of these Internet connections to transport traffic of hand-helds. Back then, these devices already contributed about 0.7 percent to the overall HTTP traffic. Further, the authors reported a sixfold increase regarding the amount of hand-held HTTP data traffic during their measurement duration, while 80 to 97 percent of all transported hand-held data stems from HTTP connections. In line with these results are more recent expectations from network equipment manufacturers published in 2011 which assume that 80 percent of wireless data traffic in cellular macro, micro, and femto cells, stemming from hand-helds, laptops, and other devices, will appear inside of buildings at residential places or office environments in the following years [111]. To account for these situations, we consider in this thesis end users who do not move during an active communication session.

Transport as well as off- and onloading of all packet-based traffic We assume that the operator of the WWAN allows to transport as well as to off- and onload all packet-based traffic, including not only web but also delay-sensitive traffic stemming from Video and VoIP applications. By this, on the one hand, we take into account that even VoIP applications such as "Skype" [112] and "Viber" [113] started to enable a transport of delay-sensitive traffic via WWANs such as 3GPP cellular networks. On the other hand, we follow approaches similar to the "Smart Offload" framework [114], where cellular operators may offload also voice traffic to WLAN hotspots.

Long-lived traffic flows As discussed above, end users tend to use their hand-held devices in home DSL/WLAN networks. Following these trends, we assume that a fraction of end users holds long traffic sessions in the order of a couple up to some tens of seconds when connected to WLANs. Actually, this was shown to appear in WLANs: Back in 2010, Gember et al. [115] compared traffic from hand-held and "non-handheld" devices in a large WLAN network on campus consisting of almost 2,000 APs. On TCP level, the authors analyzed the amount of transported payload data in the wireless downlink, which is denoted in their paper as "flow size", for both types of devices. For hand-helds, the amount of data transmitted by TCP flows shows a median of 50 KB, while the maximum lies at 630 MB. In addition, the duration of TCP connections for hand-helds is in 90 percent higher than 250 ms and in 10 percent longer than 15 s. The authors also showed that the hand-held traffic volume consists to 40 percent out of video traffic. The observed video flows on hand-helds were reported to be large in comparison to other hand-held traffic stream sizes with 80 percent of the videos being above 50 KB and 20 percent being above 1 MB. Nevertheless, video flow sizes appeared to be usually lower compared to other devices such as laptops because the streaming content is optimized for hand-held devices. Gember et al. reported 20 percent of the video streams to last longer then a second, while the median is half a second. The authors explained the duration of the video streams to be driven by the goodputs obtained in their WLANs, which have shown a median of about 2 Mbps.

4.5 Underlying Architectural Framework

Our considerations from the WLAN hotspot perspective are made within the framework of a simple architecture allowing WLAN RRM decisions in the context of the off- and onloading discussion.

4.5.1 Tasks of the Architecture

Our architecture shifts the decision processes being related to WLAN access to the managing entities of the involved 802.11 network essentially following the central controller architecture of today's WLAN enterprise networks. With this, we aim at enabling a smooth incorporation of our framework into existing architectures ranging from large enterprise WLAN networks down to single WLAN hotspot cells. As a basis for this work, we assume three main tasks that our architecture supports:

- gathering WLAN-specific measurement data from AP and end devices,
- conducting evaluations or estimations on the basis of the measured data,
- triggering actions such as handovers from the WLAN network side.

4.5.2 Architectural Components

We assume the support of all three items by means of standard-compliant schemes from the IETF, 3GPP, IEEE 802.21 as well as IEEE 802.11 standardization bodies.

For the first item, i.e., gathering WLAN-specific measurement data from end devices, we are conform to the aspects of the 802.11k/v amendments detailed in Sec. 3.7. Regarding the third item, i.e., the triggering of actions from the WLAN side, we are in accordance with the 802.21 media independent handover framework presented in Sec. 3.6.3 (compare also Fig. 3.4). There, the 802.21 *media independent handover function (MIHF)* offers a service to higher-layer MIH users. For our second item, i.e., conducting evaluations or estimations on the basis of measured data, we introduce one MIH user on network and one on device side as detailed further below. We place the first MIH user on a remote server component inside the wired infrastructure of the WLAN AP. We assume that the remote server offers calculation capabilities exceeding the abilities of plain APs by far. Finally, for a triggering of 802.11 measurements and a signaling of the measured data between a WLAN AP and an associated STA, we remain conform to the functionality of 802.11k/v with its MIB interface. From a remote server, the MIB is accessed in a standard-conform way by the *Simple Network Management Protocol (SNMP)* from IETF [116]. Basically, with its *get* and *set* commands, SNMP enables to remotely change and read out selected items in the MIB of the AP.

Figure 4.1 gives an overview about the basic entities and their interworking on the management level. We note that we closely follow the architectural option of 802.21 given in Fig. 3.4. Fig. 4.1 adds to this the standard-compliant functionality of 802.11k/v, the SNMP entities, and the MIH users residing on top of MIHF. We point out that the MIH user at network side, the RRM manager, runs the solutions presented in Chapter 6 and 7.

We shortly survey the joint functionality of our architecture. For a discussion, we group the basic entities in three categories: the resource management controller (RMC), the WLAN AP, and the multi-standard devices (MSDs) allowing wireless access to both WLAN as well as WWAN technology.

Resource Management Controller The RMC is our central entity to make handover decisions of devices with their flows residing within a WLAN hotspot. We assume that the RMC is capable to make handover decisions for a given MSD on the basis of WLAN-specific measurements, information from other heterogeneous neighbor networks, and calculations presented later in the subsequent chapters of this thesis. Overall, in 802.21 terminology, the RMC is an MIH user. As discussed in Sec. 3.7, 802.21 offers two different cases: having MIHF and MIH user either on a remote network device such as a separate server, or directly on the WLAN AP. Without loss of generality, we describe here the first, slightly more complex option. For this, our RMC consists of three parts, namely the RRM manager, the SNMP manager, and the IEEE 802.21 MIHF.

On the RMC, WLAN-related measurements for each associated MSD are triggered and requested by the SNMP protocol [116]. This data collection process is controlled by the RRM manager that steers the SNMP manager. Regarding the 802.21 services, we assume that information from heterogeneous neighbor networks or other WLAN

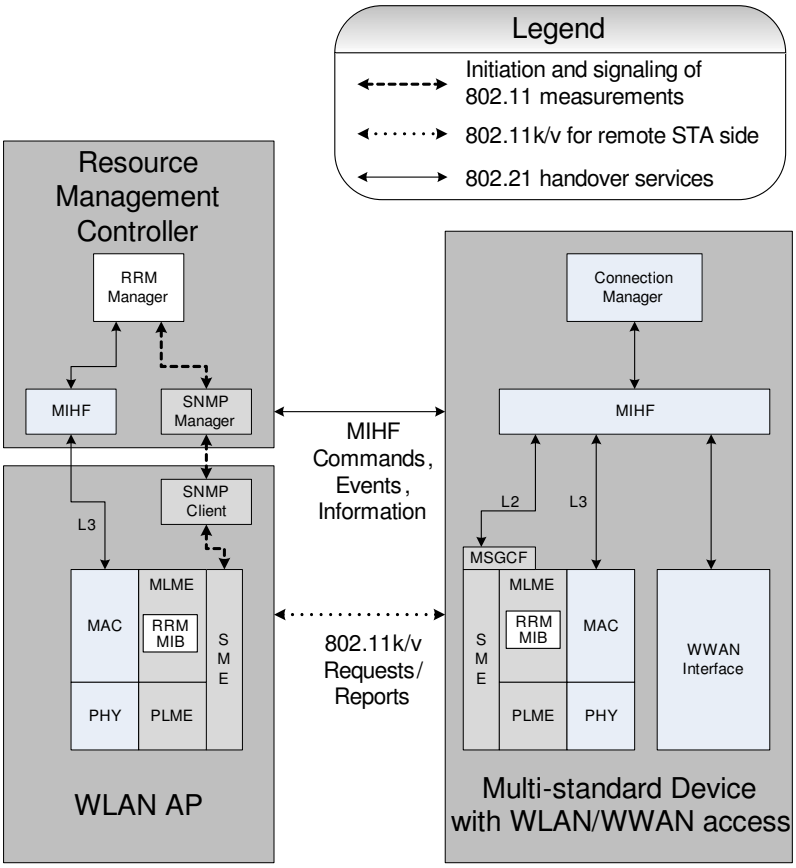


Figure 4.1: Architectural framework

networks belonging to other administrative domains may be obtained by the 802.21 standard-compliant MIHF information service. In addition, we are conform to the 802.21 MIHF command service to trigger a handover from the network side for selected MSDs currently being served inside the WLAN cell. Note that the 802.21 MIHF entity does not have any access to the 802.11 MAC management plane thus exchanging 802.21 information and the command service data using the normal path through the network stack via IP datagrams.

WLAN Access Point The second main entity, the WLAN AP, obtains radio level information from the hotspot. For this, we use the standardized schemes from the 802.11k/v amendments discussed in Sec. 3.7. As we apply network-based decisions, we aim for a data collection at the WLAN AP. Measurements are triggered via SNMP from the RMC. After the measurements, collected data is stored in the RRM MIB of the AP. This MIB is read out in our architecture via the 802.11 SME by SNMP via an SNMP client on request. On the one hand we conduct selected radio resource

measurements locally on the AP as well as remotely on MSDs. We remain conform to the 802.11k amendment with its MIB, which specifies measurements of diverse statistics on both, traffic flow and device level. Further, we make use of the 802.11k option to use proprietary metrics *only on AP-side*. We point out these metrics later in the corresponding chapters, where they are utilized. In contrast, we stick to the standard 802.11k amendment and its predefined metrics *for STAs*, both regarding the triggering of measurements on remote STA sides and the reporting of the remotely obtained data. We point out that this is an important aspect, as the 802.11 protocol stack implementation of STAs is vendor-dependent at not under control of our WLAN hotspot. As detailed in Sec. 3.7, triggering and reporting is realized by layer 2 specific request / response handshakes between the involved entities, i.e., AP and WLAN stack of the MSD. On the other hand, we are fully conform to standard WLAN network management mechanisms being introduced by the 802.11v amendment. From 802.11v we utilize the possibility of reporting device specific manufacturer information from a STA to the AP as discussed in Sec. 3.7.2. This is used throughout our estimation scheme discussed in Chapter 7.

Multi-Standard Device Finally, the MSD is the third entity in our architectural framework. Again, it is fully conform to 802.21 and the technology-dependent communication standards. It has the different NICs for WLAN and WWAN access, which in turn have the MIH_LINK_SAP interface to 802.21 MIHF (compare Fig. 3.4), translating 802.21 primitives into technology-dependent notation. For 802.11, this is covered by the *MAC state generic convergence function (MSGCF)* [6] on top of SME. Further, the MSD owns an MIHF and an MIH user, the latter being denoted as *connection manager (CM)*.

As pointed out in Sec. 3.6.3, previous works by Silva et al. [90] already introduced such a manager as an 802.21 user to control MIP in the context of vertical handovers. Again, we note that such CMs are quite common today on MSDs [117] being essentially the central instance at device level observing the QoS of on-going traffic flows, making network selection decisions (potentially on the basis of pre-defined policies of the mobile operator via ANDSF, as discussed in Sec. 3.6.1), as well as conducting the steps regarding a heterogeneous handover.

Throughout this thesis, we assume that the CM obtains commands, information and events being entirely conform to 802.21 via the interface to MIHF, both from a local as well as a remote perspective. From the local point of view, MIHF is also conform to 802.21 regarding the signaling of events from the MSD's WLAN interface such as information about low signal strength for received data frames. In addition, it is conform to the reception of handover-related signaling such as handover triggers from the corresponding MIHF entity within the RMC and forwards it to the CM.

4.5.3 Discussion of Optional Extensions

We presented above an RMC that resides on a remote server. We note that this framework fits into existing architectures ranging from large enterprise WLAN networks down to single WLAN hotspot cells as the RMC may be placed close to other different entities.

Enterprise networks usually have a centralized controller that manages a group of WLAN APs by means of *Control and Provisioning of Wireless Access Points (CAPWAP)* [118] or *Lightweight Access Point Protocol (LWAPP)* [119]. In such architectures, we suggest to simply enhance the enterprise controller by adding the RMC functionality.

For the sake of completeness, we note that one may also include the RMC functionality directly on top of the WLAN AP. In this case, the MIHF and the RRM manager directly 'sit' on top of the WLAN stack. Then, one is capable to access the 802.11 MIB locally by means of the specific operating system used by the AP. By this, our architectural framework naturally supports also single hotspots consisting of a single device only, covering the AP as well as the RMC functionality.

4.6 Thesis Scope

Considering the broad area of challenges in the context of the off- and onloading discussion, we conclude that especially the future evolvement around the WLAN hotspots will play a key role for the amount of traffic that can be shifted to such access cells by handovers. While issues like coupling and cooperation of networks, security issues as well as mobility management schemes are enhanced dominantly by standardization bodies from 3GPP, IEEE, and IETF, we expect certain critical areas of the off- and onloading discussion to be not tackled by these organizations. The majority of these aspects consists of flavors regarding schemes, their intended policies as well as their related performance metrics being usually kept vendor or operator dependent and built on top of existing solutions from standardization. As we expect the strongest gains in these directions, we put the emphasis of this thesis on the three major issues being most likely left to such vendor-specific choices. Note that we aim to focus the design of our solutions on existing as well as discussed results of the standardization bodies thus enabling a simple realization of our approaches in next generation heterogeneous networks.

The perspective of WLAN hotspots in the off- and the onloading discussion has not been discovered well, although a justified radio resource management in these cells may lead to the most significant gains regarding the amount of offloaded traffic. Specifically, this thesis focusses on three most important key enablers in different dimensions for a fine-grained support of traffic onloading in WLANs. Throughout this work, we propose novel schemes in the following areas, namely regarding

1. an opportunistic preparation of single-radio handovers,
2. the handover candidate selection in WLANs, and
3. the WLAN link data rate estimation.

Opportunistic Preparation of Single-Radio Handovers The first part presents a solution for a preparation of onloading handovers. From the perspective of the architectural framework in Sec. 4.5.2, our solution solely applies on multi-standard device side.

The motivation for our contribution is as follows. Loading up a given WLAN hotspot may also require to shift selected traffic away, as further detailed below in our second contribution. Moving such traffic away from a given hotspot can be conducted in our scenario both on a homogeneous as well as a heterogeneous network level. Multi-standard devices usually allow a parallel usage of WLAN as well as WWAN technology, if for both a separate NIC resides inside the MSD. Thus, a handover from a WLAN hotspot to a WWAN can be conducted in a soft fashion by using both heterogeneous links at the same time. In contrast, both WLAN and WWAN technology can be incorporated on a multi-mode NIC because of cost, energy, and size issues as discussed in Sec. 4.3.3. Then, if both technologies share certain pieces of hardware, such as a transceiver chain, no parallel usage of both is possible. Under such hardware constraints we propose a solution to handle cases where single-radio onloading handovers may appear.

In our selected scenario in Sec. 4.4, a single-radio handover essentially happens in two cases in the context of onloading or shifting traffic away from a given hotspot: either when we aim to shift some traffic from one WLAN cell to another (homogeneous case), or if we hand over traffic from a WLAN cell to a complementary technology that requires access to some shared hardware parts on the end-user devices being already occupied for WLAN transmissions (heterogeneous case). For both cases, we drive the fundamental approach to pause ongoing (WLAN) communication for extremely short but frequent time spans. Thereby, the time span of a pause may just last a couple of milliseconds. In each of these time spans, we switch to a selected wireless channel and conduct a handover preparation to another network. For this, we subdivide the handover preparation into small steps, such that each can be applied in a WLAN communication pause separately. While, the general solution idea remains the same for both, the homogeneous as well as the heterogeneous case, the exact approach with its switching policy differs for each as a result of different MAC schemes for the selected technologies and their timing behavior. We analyzed and evaluated our approach regarding two selected application examples covering homogeneous WLANs as well as heterogeneous WLAN/WiMAX networks. Thereby, our results identify the limits regarding timing dependencies and background load as well as the range of the signaling overhead induced by our approach.

Handover Candidate Selection in WLANs Sec. 4.3.4 discussed that a resource-optimal allocation of traffic flows to a WLAN hotspot remains challenging, even in the presence of admission control schemes. What has been neglected so far in the off- and onloading context, is an identification of devices that are not suitable for a certain WLAN hotspot. The term unsuitability thereby refers to traffic streams with negative impact on the capacity of the wireless cell. Actually, today's vertical handover mechanisms are in the majority user-centric approaches which have been designed to handle the mobility of end users. Yet, these mobility-oriented schemes lack a proper support from the WLAN level giving feedback about the suitability of each device, thus allowing to cover also resource-aware onloading handovers.

On 802.11 MAC and PHY level, a whole 'zoo' of technological parameters is available, e.g., measures of the received signal strength, the required number of retransmis-

sions, the selected MCSs for transmissions, or contention level parameters, reflecting the operational point of the WLAN hotspot. From these parameters, one can easily obtain the time that the wireless channel is occupied by a given traffic stream, commonly denoted as the 'airtime'. Nevertheless, it not easy to derive from this which traffic stream out of many is actually negatively influencing the whole cell in terms of capacity. However, for a maximum loading of the cell, a WLAN operator is required to identify unsuitable traffic streams that should be selected for an onloading, i.e., a heterogeneous handover back to the WWAN. We refer to this as the *handover candidate selection*.

For a clear distinction between suitable and unsuitable traffic streams, this thesis firstly presents a performance metric aiming at an efficiency evaluation in terms of the occupied resources, i.e., the airtime. Our performance metric easily aggregates selected WLAN MAC and PHY parameters into a unique measure, such that it covers detailed technological insights with only low computational effort. After defining our metric, we first demonstrate its usability. Second, we evaluated our metric in a scheme for the selection of handover candidates in a WLAN cell to be onloaded to the WWAN by comparing the gains of our approach with the classical RSSI-based as well as simple random decisions. For different operational points of the WLAN, our results show that our selection approach outperforms the classical schemes primarily in dense settings.

In the architecture defined in Sec. 4.5.2 and shown in Fig. 4.1, both the AP and the WLAN NIC of the device are assumed to conduct the measurements of selected 802.11 parameters. Via 802.11k, the AP collects the measurements from the device. In regular time intervals in the order of seconds, the RRM manager triggers a collection of the measured parameters from AP and conducts a calculation of our performance metric for selected MSDs. Finally, on the basis of this metric, the RRM manager makes handover decisions that are signaled to the MSD via the 802.21 framework.

WLAN Link Data Rate Estimation During offloading decisions, current network selection schemes are not able to predict the behavior of a device regarding its resource consumption in a WLAN cell well. Network selections based on accurate predictions regarding the behavior of offloaded traffic in a WLAN cell would be the key solution to greatly improve such schemes. Unfortunately, such a prediction would be most likely very imprecise today as each WLAN device includes vendor-specific algorithms whose effect on a given WLAN hotspot are unknown. A well-known example for such a vendor-specific behavior are the rate adaptation schemes with which each WLAN NIC adapts the MCSs for its wireless transmissions. Aiming to avoid devices which transmit only with MCSs resulting in low link data rates, we envision that a device first associates with a target WLAN cell being determined by classical, rather heuristics-based network selection schemes. Then however, we observe the first transmissions of this device in WLAN afterwards. These transmissions could be, for example, related to the signaling of a handover or may even include the first data transmissions on the WLAN link. On the basis of these observations, we make an estimation whether the newly associated device will be suitable in the WLAN cell on a long-term scale regarding its link data rate selection behavior. We emphasize that we rely our estimation on devices with stationary

traffic, where we have a priori knowledge about the traffic pattern. In other words, we do *not* predict the actual traffic pattern of a device, but focus on the link data rate adaptation behavior for this given type traffic. To this, we refer to as the WLAN link data rate estimation. We expect predictive decisions about WLAN devices to further improve the gains for offloaded traffic as we aim to avoid having inefficient traffic streams that waste WLAN resources by selecting only robust MCSs. How to realize such estimates under the assumption of diverse, vendor-specific behavior among different devices has been probably the most challenging point in our studies.

In this context, we present our innovative *link data rate estimation scheme 'DARA'* for WLAN hotspots. The core mechanism of DARA estimates the link data rate selection of a WLAN end-user device just by observing its behavior on short time scales—without having any knowledge about the applied rate adaptation algorithm. For the estimation, we present the selected machine learning approach used by DARA and demonstrate its potential using both simulated WLAN configurations as well as measurements from a WLAN setup in an on-campus office environment. We studied DARA's estimation accuracy for the selected rate adaptation schemes adaptive auto rate fallback (AARF) / AMRR and Minstrel with data from simulations as well as measurements. Finally, we utilized DARA's estimates as a basis for selecting 'suitable' devices and to onload 'unsuitable' devices back to the WWAN instead.

From the architectural perspective given in Sec. 4.5.2, we assume again that the AP and the WLAN NIC of the device conduct measurements of selected 802.11 parameters. Via 802.11k/v, the AP collects not only the measurements, but also obtains vendor-specific information from the device. The RRM manager triggers a collection of the data from the AP and, on this basis, runs the estimation scheme. Finally, the estimation results are used by the RRM manager to make handover decisions, which are signaled again to the MSD via the 802.21 framework.

Chapter 5

Opportunistic Preparation of Single-Radio Handovers

As detailed in the preceding chapters, dynamic decisions about the access technology to be used seem to be a promising approach to overcome throughput bottlenecks. While quite a lot of research is devoted to a proper selection of the access technologies as described in Sec. 3.4.3, it is frequently taken for granted that the switching itself, i.e., an enforced handover from one technology to another, is somehow easily done. Switching from one technology to another while supporting a continuous communication is, however, by far not trivial. A *preparative process of a handover* on layer 2 includes the neighbor network discovery, the decision about the availability of a certain technology for a given type of end-user traffic, the selection of a proper access point, and finally the completion of the association procedure. The sum of all these steps can take a significant amount of time. Thus, it is hardly recommendable to enforce a 'hard' handover, i.e., breaking the ongoing communication association before assuring that another, better connectivity is available. Realizing a soft handover with service continuity is restricted, however, by the applied communication hardware. Even if multiple network interface cards (NICs), each supporting one of the technologies, were available, their parallel usage would be avoided in order to economize energy. In reality nowadays, there is a strong trend to limit the number of separate NICs within a single device due to space and cost issues. As a result, multi-standard NICs become more and more a solution of choice. We refer to such NICs as multi-mode radios if they can support multiple technologies, but at a given time only access to one wireless technology over a specific transceiver chain may be possible. As introduced in Sec. 4.3.3, we refer to a handover over such a NIC as a *single-radio* handover.

The preparation of such a handover is not only an issue for a support of different wireless technologies over a multi-mode radio, but also appears in a similar analogy in the case of homogeneous handovers, where the end-user device is just equipped with a single NIC of a given technology. A typical example is the mature WLAN technology, which is known for its lengthy neighbor network discovery. Nevertheless, in the past, network discovery was a rather seldom occurrence. Over the few last years, with the

increasing popularity and density of WLAN hotspots, this has been changing: WLAN devices have not only been faced with fluctuating channels as result of interference, but also have experienced increasing temporary high-load situations strongly degrading the throughput of each.

In the context of onloading, we aim to shift selected devices from a WLAN hotspot to other heterogeneous wireless accesses in our reference scenario given in Sec. 4.4. However, devices may not only conduct a handover to the WWAN cell, but can be also shifted to a second WLAN hotspot as discussed in Sec. 4.3.3 and 4.6. As a basis for this chapter, we extend the reference scenario. There, WWAN comprises of mobile WiMAX technology (802.16e). Further, we have exactly two WLAN cells which have a significant overlapping coverage area. Both APs of the WLAN hotspots operate on non-overlapping channels and transmit beacons to announce their existence at regular time intervals as defined by the 802.11 standard. Thereby, the APs do not coordinate or synchronize the process of sending the beacon frames. In the area covered by both hotspots, one multi-standard device with a multi-mode radio for WLAN/WiMAX is present. This device is associated with the first WLAN hotspot over which it runs a delay-constrained service. Throughout this chapter, we consider that the device runs a VoIP call with a peer on the Internet. In the following, we refer to this device also as the *opportunistic STA*, as it applies our opportunistic approach. By the term *opportunistic*, we emphasize that our solution is applicable and adaptable to scenarios both for homogeneous as well as heterogeneous onloading handover cases.

Further, we assume that the connection manager (CM) of the opportunistic device initiates specific handover preparation actions detailed later. The CM may have either decided on its own, e.g., based on QoS measures for the VoIP stream, or may have been triggered by the RMC, residing in the backplane of the first hotspot. In this chapter, we assume the following actions. The multi-mode device conducts a handover preparation either a) for an alternative WLAN hotspot or b) for the WWAN (mobile WiMAX). Further, we assume that the CM by means of the 802.21 information service is provided with the specific channel information on which the alternative WLAN or WWAN access is operating. This information has been gathered before from the RMC. Independent of the considered technology to which the handover preparation takes places, we denote its channel as the *secondary communication channel* in the following. In contrast, we refer to the *primary communication channel* as the channel of the first WLAN hotspot, in which the regular data transmissions of the considered device takes place.

We formulate the following design goals for the handover preparation over the multi-mode NIC whereby we aim to support in parallel an on-going, delay-constrained traffic stream via the first hotspot:

- The approach should enable real-time communication with small packet inter-arrival times and hard QoS constraints requiring low packet loss and relative small extra delay at MAC. Such applications include, e.g., VoIP traffic, which is a typical representative for such pattern (e.g., G.711-coded speech [120] with 20 ms inter-packet generation time).

- We aim to adapt our approach to a timing pattern on the secondary communication channel. Thus, the timing of the approach should not solely depend on the inter-packet generation time of the on-going traffic stream. Instead we aim to enable a switching between primary and secondary communication channels even on time spans below inter-packet generation times.

- To scale with the number of end devices, the approach shall not (unproductively) affect any communication on the secondary communication channels, e.g., by active scans in WLANs.

- A feasible solution in real hardware can be only built upon mechanisms that are standard compliant. Most wireless technologies incorporate means for power saving mechanisms, even allowing to pause communication on small time scales in the order of tens of milliseconds. Accordingly, we rely our approach on selected power management mechanisms to support a vast group of technologies.

This chapter first discusses related work and presents a generic design aiming to fulfill a handover preparation on multi-mode NIC architectures. Second, we evaluate the presented approach by focussing on two specific examples. The first considers the network discovery for homogeneous WLAN networks while the second focusses on a handover preparation for a heterogeneous WLAN/WiMAX network. Further, we present the evaluation methodology for both examples and present the corresponding results afterwards. We point out that the work regarding the homogeneous WLAN network discovery was presented already in [5, 121], while the handover preparation in a heterogeneous WLAN/WiMAX network was published in [122].

5.1 Existing Support for Single-Radio Handovers

This section gives a survey about existing approaches for single-radio handovers. We start with a review about homogeneous WLAN handovers and related enhancements. Afterwards we consider previous works regarding heterogeneous handovers on single-radio architectures. We end with a summary of the relevant aspects for this thesis and discuss related challenges.

5.1.1 Homogeneous WLANs

Mishra et al. [123] analyzed the duration of handovers for STAs migrating between different WLAN APs. First, the authors showed that the total handover duration can result in an interruption of the layer-2 connectivity on average up to 400 ms, which is by magnitudes far above QoS requirements of delay-constrained traffic such as VoIP. Second, Mishra et al. further detailed the duration of each handover phase, i.e., neighbor discovery, authentication, and (re-)association (c.f. Sec. 2.4 regarding the WLAN network entry). Their measurements identified that the neighbor discovery phase contributes to more than 90 percent to the overall handover duration, even though the authors apply

802.11 active scanning. Further, more than 80 percent of the frames during a handover are transmitted because the STA actively scans the WLAN channels.

Certain approaches [124,125] reduced the duration of neighbor discovery for STAs by coordinating the beacon transmissions of all WLAN APs in a given channel. In WLAN networks offering such a support, STAs knowing the schedule for beacon transmissions in each channel need to conduct short passive scans only during dedicated time frames. As these specific solutions assume a support from the WLAN-network side regarding a coordination of surrounding APs, from our perspective, they are not viable for our intended single-radio handovers which shall operate also in (uncoordinated) hotspot scenarios.

In contrast, other schemes focussed more on a reduction of scanning times without requiring means from the WLAN network or AP side. While some authors adapted the STA parameters for the active scanning regarding the load level in the scanned channel [126], others proposed to use "proactive" scans [127] where STAs alternate active scans on single WLAN channels with on-going communication in time far before a handover is conducted. Teng et al. [128], among others, suggested to consider during active scans not only probe response frames, but also other up- and downlink traffic in which information about APs is conveyed. Further, regarding neighbor network discovery schemes conducting also a selection of channels to be scanned, we refer the reader to the survey of Pack et al. [129].

Alternating short scans on selected WLAN channels with on-going data communication in time was shown to reduce the duration of interrupted layer-2 connectivity and related packet losses [130]. Nevertheless packet losses still appear if a STA is scanning another channel, while its primary AP starts a data transmission. Tackling this problem for elastic traffic, Chandra et al. [131] presented an approach for a device being only equipped with a single WLAN NIC back in 2004. The authors introduced multiple "virtual adapters over a single WLAN card", thus allowing a "multiplexing" of WLAN associations with different hotspots in time [131]. This approach relies on the basic 802.11 power save feature to handle the absence of the device from one BSS, while being active in other hotspots. However, with a basic 802.11 power save, the interruption of connectivity for a NIC from its BSS may last multiples of the WLAN beacon interval. Thus, this approach works for elastic web traffic but may fall short together with delay-constrained voice or video streaming applications.

From the perspective to improve the energy consumption of a WLAN STA during a communication session with delay-constrained VoIP traffic, the work of Chen et al. [132] applied the standard-compliant, basic WLAN power save mechanisms. By carefully configuring the 802.11 power save with extensions from 802.11e, the authors showed that these mechanisms are indeed able to support VoIP traffic. Borrowing some of this functionality, more recent works extended the basic idea of Chandra et al. [131] "multiplexing" WLAN associations with multiple APs on time scales below a beacon interval, enabling improved throughputs [133–135] as well as handovers [133] for TCP traffic. Kandula et al. [135] showed with a practical implementation for off-the-shelf hardware that a WLAN STA associated with multiple APs can be switched from one to another on average in 3 ms including standard power save mechanisms.

Further, on the basis of reference [132], Tsao et al. [136] presented an approach of "soft" WLAN handovers for VoIP traffic. There, the authors proposed, for a WLAN STA with a single NIC, to use two MAC addresses, claiming to allow for two simultaneous associations with different APs. Further, their scheme "DualMAC" subdivides the steps of handover mechanisms in time. Aiming at a support of on-going VoIP traffic with one AP, their approach utilizes the time between two consecutive VoIP packets on STA side to conduct handover preparation steps with a second AP. To avoid a loss of VoIP packets, the STA pauses its communication with its original AP by means of 802.11 power save mechanisms. Thereby, Tsao et al. based the timing exactly on the generation rate of uplink VoIP packets of the STA. The authors noted that they can support a minimum of 20 ms time gaps between two consecutive frames to allow for a completion of an active scan on a single WLAN channel. By this, Tsao et al. showed to introduce additional VoIP packet delays of up to 50 ms in their simulations, primarily resulting out of the active scanning waiting for probe response frames.

5.1.2 Recent Trends in Multi-Mode Radios

In order to enable heterogeneous wireless access anywhere and anytime with just a limited number of NICs within a device, a hybrid approach is common today. Multi-mode, reconfigurable radios are able to do MAC as well as some PHY (base-band) processing in pure software, but still apply transceiver chains with specific analog parts (amplifier, filter) that are specifically designed and adopted to their purposes and frequency bands (e.g. *joint* WLAN/WiMAX and 2G/UMTS/LTE transceiver chains) [137]. Vendors like Infineon or Intel [138, 139], for example, have been developing dual-mode WLAN/WiMAX transceiver chips that have been incorporated into today's end-user devices. However, such highly integrated multi-mode radios support at a given time only access to one wireless technology over a specific transceiver chain, such that no simultaneous, parallel usage of the same chain for another technology is possible [140].

A trivial solution may simply switch the transceiver chain from one access technology to another, imposing a hard vertical handover. Choi et al. [141] proposed a more sophisticated heterogeneous handover scheme, for multiple NICs, in which it is possible to have only a single NIC active during each time instance. This does not only decrease energy expenditures of a device but also reduces the amount of occupied resources from the networks. The authors utilized the "mutual silence periods" of ON/OFF traffic patterns, such as VoIP, for the execution of handovers. Both, hard vertical handovers as well as Choi's scheme, may be promising approaches for technologies with quick layer 2 associations. In contrast, they may suffer in case of handovers to a technology, e.g., such as WiMAX, which is known to have a very lengthy network entry [142] thus exceeding the length of Choi's "mutual silence periods" by magnitudes.

Most existing works regarding an operation of multi-mode radios considered WLAN/ WiMAX, where the device is already associated with both types of networks. WiMAX is thereby usually operated in the *time division duplex (TDD)* mode, in which a strict timing pattern regulates medium access such that downlink and uplink transmissions are conducted in separate time frames, also denoted as *down- and uplink subframes*. One

down- and one uplink subframe together form a *WiMAX time frame*, which has usually a duration of 5 ms for mobile WiMAX [143]. The scheduling of transmissions in each frame is coordinated centrally by the *base station (BS)*, which is comparable to an AP in 802.11 networks. Yang et al. [140] proposed for a WLAN/WiMAX multi-mode device to use the WiMAX uplink subframes for WLAN access if the device does not have to transmit a packet within WiMAX. Further, the authors propose a scheduling that aggregates WiMAX uplink traffic maximizing the number of free uplink subframes for the access in WLAN. Considering multiple multi-mode devices connecting to the same WLAN, the authors further provide a WiMAX schedule that prevents all devices to return and initiate transmissions in the WLAN hotspot. Other work [144, 145] suggested to use power save mechanisms in WiMAX, enabling to switch to other technologies such as WLAN or Bluetooth during the sleep times. However, we note that for the setup of a sleep discipline in a WiMAX network, the device must be already associated with a WiMAX BS. Thus, the power save mechanisms do not help if a device still has to undergo the network entry in WiMAX.

5.1.3 Summary and Discussion

While for homogeneous WLAN handovers, approaches from existing work propose to alternate data communication and handover preparation steps in time, they still fall short regarding several aspects that we discuss in the following.

One issue with previous works lies in the dominant usage of active scanning to reduce the duration of single neighbor discovery steps. However, active scanning imposes significant flaws which prevents its usage in many existing and upcoming real-life scenarios. First, active scanning fails to comply with requirements of certain regulation bodies such as the Federal Communications Commission (FCC) and the European Telecommunications Standards Institute (ETSI). In the European Union, ETSI effectively prevents active scanning for 802.11a/n in the frequency range from 5250 to 5350 MHz and from 5470 to 5725 MHz in order to protect radar equipment operating in the same bands [146]. With the channels for 802.11a given in [3], we conclude that in 15 out of 19 5 GHz WLAN channels, active scanning is prohibited. Further, the 802.11af task group, aiming to adapt 802.11 for an operation in TV whitespace channels, limits the application of active scans to cases where STAs obtained a "white space map" before [147], essentially giving information about useable channels in time, frequency, and space. Since information of such "white space maps" is restricted to a limited range in space, e.g., the coverage area of a BSS, it remains questionable to what extend mobile devices can be supported. Thus, existing single-radio handover schemes relying on active scanning will definitely not cover the full range of scenarios in upcoming WLAN channels.

The second severe problem with active scanning appears due to its overhead in terms of the time that it occupies a WLAN channel. To a broadcast-addressed probe request frame, all APs in the vicinity respond with an own probe response frame. Thus, the load on a scanned channel directly depends on the AP density in the vicinity of a scanning STA. In addition, the more STAs in parallel conduct an active scanning, the further increases the occupied time of the channel. Raghavendra et al. [148] evaluated the

relation of the probe traffic and the population of WLAN STAs from measurement traces gathered at an IETF meeting. Their findings identified that not only the active scan traffic increases with the number of STAs, but also that losses of requests and responses are affected as the load in the network grows. This in turn induces additional scan traffic. From the perspective of a large Japanese telecom operator, Yunoki et al. [149] argued in a contribution to the 802.11ai task group that specifically for cellular traffic offloading to WLANs, a support of crowded scenarios, i.e., a high density of WLAN STAs in a certain area, would be advantageous. The authors presented WLAN traffic statistics recorded in an environment where around 23 STAs simultaneously conducted active scans on a channel with 4 to 5 present APs. From the measurements, Yunoki et al. calculated that probe request and response frames together occupy more than 18 percent of the channel in time. In a follow-up contribution [150], the authors interpolated the occupied time of the channel to be above 75 percent for 100 WLAN STAs conducting active scans.

As a result of the above aspects, we conclude that actively scanning WLAN channels for a preparation of homogenous single-radio handovers is not a viable option in the context of this thesis. Instead, we aim to rely on passive scanning. However, traditional passive scanning stays on a WLAN channel up to multiples of a beacon interval usually being around 100 ms. Thus, we aim to extend the approaches of related work by splitting a passive scan on a *single* WLAN channel into multiple, small sub steps. Thereby, we aim to alternate on-going delay-constrained communication with the scanning steps in time.

Although the work of Tsao et al. [136] already aimed to apply an (active) scan in a minimum time gap of 20 ms between two VoIP packets, it is not viable for our purposes. First, it uses active scanning which, besides the drawbacks discussed above, introduces significant additional delays depending on the number of surrounding APs and the background load in the channel. From their results showing timely arrivals of VoIP packets before and after a handover, we conclude that their considered WLAN setup seems to have been in a low load state. Still under these circumstances, the authors reported additional delays up to 50 ms. As result, even in the low load state, they are bounded to minimum interruptions of 20 ms for a scan attempt. In addition and even worse, their scheme is triggered by the pseudo-regular uplink VoIP transmissions which in turn govern the whole timing of the neighbor discovery. In contrast, we aim for a timing that is *independent* of the generation rate of uplink traffic and further allows to scan even in much smaller slices of time down to only some milliseconds. This does in principle not only allow for delay-constrained, non-periodic traffic with higher packet generation rates in the uplink such as video, but also enables us to adapt the timing of our approach to requirements of the system that we aim to detect by scanning.

Finally from the perspective of heterogeneous multi-mode NICs, we conclude that handovers towards a technology such as WiMAX, having a lengthy network entry, impose significant issues. Related work for associated multi-mode devices suggest to use power save mechanisms in WiMAX to pause communication and switch to WLAN during the sleep times. This power save mechanism, however, is not an option for a device that still has to undergo the network entry in WiMAX. Another promising approach from Yang et al. [140] suggested to switch to WiMAX in the downlink and to WLAN during a free

uplink subframe. Such a scheme, however, requires to adapt the WLAN access strictly to the timing of WiMAX. As this scheme does not rely on power save mechanisms in WLANs, it offers no protection against packet loss.

5.2 The Opportunistic Approach

Our basic concept is driven by the paradigm that seamless connectivity perceived by the end user may still allow for interruptions on layer 2, if the duration and the frequency of these 'layer 2 breaks' is not violating the specific QoS constraints of the application in use. To facilitate understanding, we shortly discuss the general concept taking an example from everyday life: the procedure of driving a car. There, the driver takes care most of the time about what is happening ahead. However, with a certain frequency, the driver also takes very small periodic breaks to observe the scene behind by having a look at the rear mirrors. In our analogy, the surveillance of the scene ahead is nothing else than our on-going wireless communication. In contrast, the short periodic breaks for the usage of the mirrors are the 'layer 2 breaks' being used for handover preparation steps in another communication channel either belonging to the same, homogeneous network or to a second wireless technology.

5.2.1 Range of Design Options

Now, staying a moment in our car example, we have three degrees of freedom: the exact *point in time* at which the driver looks in the mirror, the *duration* for the mirror view, and the *periodicity or frequency* with which the driver repeats his pattern. The first issue can be influenced by specific events, such as hearing alarm horns, recognizing a change of traffic lights or the appearance of some other event in front of the car, which the driver classifies as high-priority event. Such events let the driver further observe the scene ahead, thus postponing the usage of the mirrors. Further, the duration for the mirror view and the frequency for switching back and forth depends on the specific target pattern to be analyzed via the mirror and the actual situation which the driver perceives himself. In other words, a driver adapts the duration and the frequency of his mirror views to a given situation. For example, both a high frequency and a low duration may be chosen, if our driver travels on a high-way and observes a fast vehicle appearing from the rear coming rapidly close in a short period of time. Instead, both a rather low frequency and a long duration may appear if our driver stopped at traffic lights and, just out of curiosity, monitors a specific person sitting in a car behind.

Coming back to the handover preparation on a given multi-mode device, similar to our car analogy, also the point in time to switch to the secondary communication channel may be influenced by certain events. Depending on the specific design, such events can be assigned a higher priority than the switch to the secondary channel. As such a high-priority event, we consider in our work pending, time-critical data transmissions stemming from delay-constrained services such as VoIP. We refer to this as the *prioritization of on-going communication over the handover preparation.*

In contrast, our second design option trades off the duration on the secondary channel and the frequency for switching back and forth between the primary and secondary channel. Again, on the primary communication channel, the timing requirements and the periodicity of the traffic pattern to be transported plays an important role. However, also the technological MAC behavior on the secondary channel may further impose specific timing requirements, thus influencing the duration and the switching frequency. If high priority is given to the latter two aspects, thus adapting strictly to the pattern on the secondary communication channel, we refer to it as to *favor the handover preparation over the on-going communication*.

In summary, from the illustrative car example discussed above, we derived two different design options for our opportunistic approach, which alternates on-going wireless communication as well as handover preparation steps on small time scales. Conceptually, we aim to cover both design options such that we are able to

1. prioritize the on-going communication over the handover preparation, or
2. favor the handover preparation over the on-going communication.

The actual selection of a specific flavor broadly depends on the application scope in our scenario. When a fast and time-constrained handover is required, e.g., as a result of sudden and strong impairments in the end-to-end QoS, most likely the second solution may be of interest. In contrast, the first option may be favored, e.g., for a regular background search of alternative channels. We note that both directions may appear in the context of onloading handovers.

5.2.2 Selected WLAN Application Cases

Above, we introduced the basic idea of the opportunistic approach without relying on a specific technology. However, for our reference scenario and the discussion of shifting selected devices with their traffic away from our first WLAN hotspot, we introduce two concrete application cases. From this, we derive a precise design for WLAN devices essentially covering both design options from the preceding subsection. Again, we aim to support a handover preparation from our first WLAN hotspot either to a) the second WLAN cell or b) to the WWAN (mobile WiMAX).

We formulate both application cases such that they cover both dimensions of the design options. In the first application case, we prioritize on-going WLAN communication in the first WLAN hotspot over a handover preparation to the second WLAN cell. For this, we consider the transport of data from a real-time VoIP application together with a periodic network discovery in homogeneous WLANs. We focus specifically on the network discovery process, as it is known to be lengthy so that it originally leads to a violation of QoS constraints for VoIP services. However, it is quite common for WLAN-commodity hardware that idle WLAN devices, not currently transporting any traffic streams, regularly scan their environment. How to conduct such a regular scan without actually harming the QoS of an on-going VoIP call is the key question for our first application scenario. For this, we first apply the general idea of the opportunistic

approach for a regular neighbor discovery in homogeneous WLANs denoted as *opportunistic scanning* in the following.

Afterwards, in our second application case, we consider our multi-standard device, amongst others equipped with a WLAN/WiMAX multi-mode NIC, to conduct a handover preparation in WiMAX. Again, the end-user device has an on-going VoIP call via the first WLAN hotspot. For a support of a timely handover, we favor the preparation for a handover from WLAN to WiMAX over the on-going VoIP communication. Although we put the emphasis on the timing priority of WiMAX, we show that for a broad range of factors, the VoIP QoS may still remain on an acceptable level for the end user. We denote this specific application of our approach as the *opportunistic preparation of handovers from WLAN to WiMAX*.

5.2.3 Reuse of the 802.11 Power Management

Both design options have in common that we consider an on-going VoIP communication session of an end-user device via WLAN. We selected VoIP as an example of delay-constrained traffic to enable a comparison with related work. As discussed above, the device should conduct short layer 2 breaks for handover preparation steps. However, VoIP traffic is known to be very sensitive to packet loss. Although the interruptions of the layer 2 connectivity are intended to be short (a couple up to tens of milliseconds), we are required to use a mechanism avoiding packet loss as a result of the temporary absence of the WLAN device from its primary communication channel.

As described in Sec. 5.1.1, various approaches have been already utilizing 802.11 power management functions to avoid packet loss. Back in Sec. 2.8, we already discussed the basic 802.11 power management functionality. However, as pointed out in Sec. 2.8.3, the plain power management schemes introduce strong additional delays for data traffic. Thus, we cannot rely on these basic schemes solely. Further, we cannot simply use the WLAN power management as it is was done in previous work by Tsao et al. [136] for a handover preparation, where data packets were used for the power save signaling. Remember that we just want to have an interruption of the WLAN connectivity for the opportunistic STA in the order of a few up to just some tens of milliseconds that can be even below the generation rate of the uplink traffic. For this, we apply the 802.11 power management, whereby we rely the signaling on null-data MAC frames if no data packet transmission is available. This has been widely used before on 802.11 NICs [151,152], e.g., to adapt the sleep discipline to the traffic pattern, as it avoids the usage of PS polls (where the PS STA stays in PS mode the whole time and explicitly triggers the transmission of downlink frames only after receiving an indication flag in a beacon frame). Thus, also in the context of WLAN devices handling multiple associations with different APs, e.g., in [133–135], the process of switching between APs includes a power-save signaling by means of null-data packets.

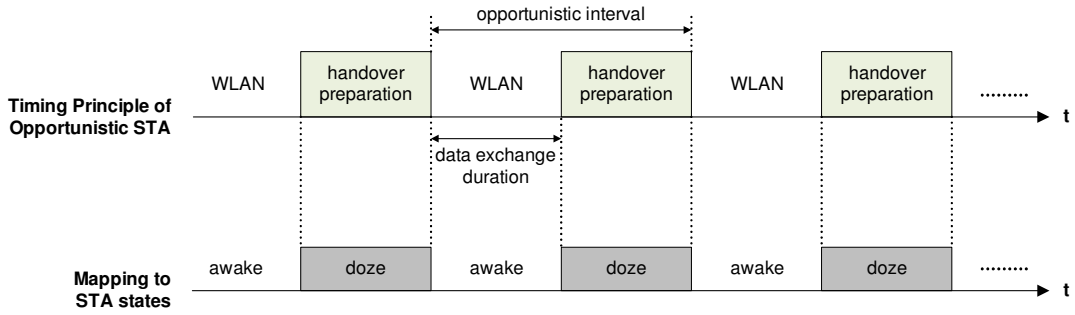


Figure 5.1: Opportunistic approach: timing principle and mapping to WLAN STA states

5.3 Design of a Holistic Scheme for WLAN Devices

As described in Sec. 5.2, to enable real-time communication with small packet inter-arrival times and hard QoS constraints, we alternate phases of data transmissions in WLAN with handover preparation steps on time scales of a couple to some tens of milliseconds. We introduce the notion of an *opportunistic interval* which includes both a time span for the exchange of data frames and a subsequent slot for the handover preparation. The basic timing principle of an opportunistic STA is shown in the upper part of Fig. 5.1. Back in Sec. 5.2.3, we argued that we aim to avoid packet losses for the on-going data transmissions. For this, we utilize the 802.11 power management in the data exchange time slot. The lower part of Fig. 5.1 maps the two phases of our opportunistic interval to the 802.11 power management STA states for the primary communication channel that we reuse for our scheme. Generally, the 802.11 standard intends a WLAN STA in power save (PS) mode, also denoted as PS STA, to switch off its transceiver chain while being in doze state. However, as discussed in Sec. 5.1.1, many approaches use this time span for other actions. In a similar way, the opportunistic STA uses this time span for handover preparation steps instead.

The opportunistic STA initiates our approach by entering the PS mode, as detailed in Sec. 2.8.2, by transmitting a (null) data frame with the power management (PM) bit set to one. Then, in a handover preparation slot, we conduct the selected means depending on the specific application case detailed further below. Subsequently, in the data exchange phase, our STA re-activates the WLAN link by means of 802.11 power management signaling. Although we use 802.11 standard compliant approaches, even the plain 802.11 PS procedures allow for different combinations, which we discuss in the following separately for each design option that we introduced in Section 5.2.1, which are namely 1) to favor the handover preparation over the on-going communication, and 2) to prioritize the on-going communication over the handover preparation. The major challenge for us thereby was to enable a support in principle of both options with one holistic scheme, such that we just imply minimal changes to existing and well-working 802.11 power save implementations by keeping standard conformity. Overall, we have

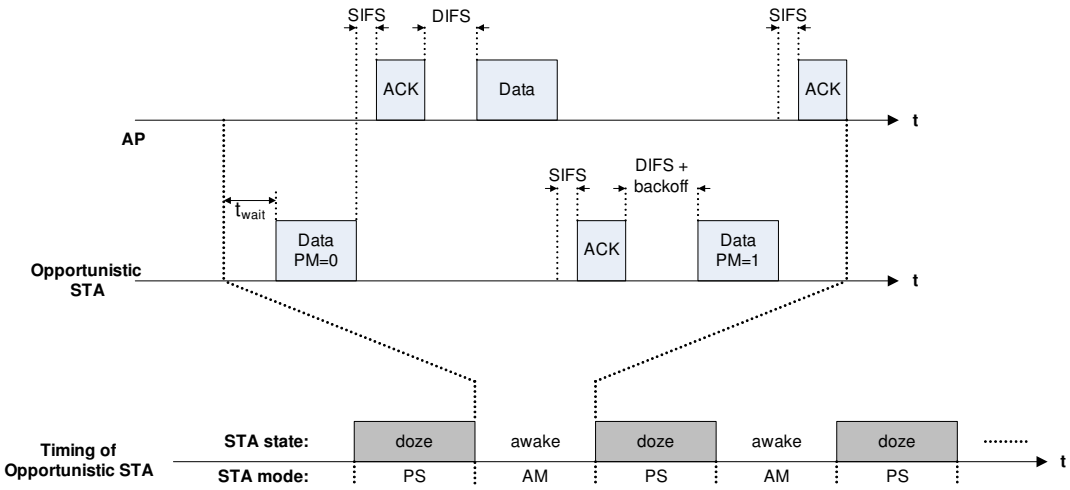


Figure 5.2: WLAN PS signaling for favoring the handover preparation

been aiming to stick to the most simple 802.11 power management features allowing our holistic solution to be applied over a broad range of scenarios and requirements.

5.3.1 Favoring the Handover Preparation

Giving priority to handover preparation steps may require a high flexibility of the opportunistic STA to adapt to a given schedule. For our selected application case, we conduct a network entry to WiMAX in the handover preparation slots. The detailed timing highly depends on the selected technology on the second communication channel. We describe these aspects regarding WiMAX in detail in Sec. 5.7.1. Nevertheless, we point out that we aim to strictly adhere to the timing of the handover preparation time slices. In case that the data exchange in WLAN takes less time, we switch to the second communication channel earlier. In contrast, if the PS signaling with the data exchange is going to be not completed on time, we basically have two choices: either to switch shortly before its start to the handover preparation phase on time, or to skip this phase. By design, we base our choice on the context of the handover preparation process in the second communication channel, as detailed later in Sec. 5.7.1.

From the WLAN side, this adoption to strict timing priorities (e.g., the frame structure of a heterogeneous technology such as WiMAX) limits us in the design space, as the opportunistic STA needs to keep control about the whole timing process. However, the plain 802.11 PS mode approach with PS polls (compare Sec. 2.10) and even sophisticated enhancements of 802.11e[1] require a STA to handoff some parts of this timing control to

[1]The 802.11e amendment [20] further specifies *automatic power save delivery (APSD)* procedures defining specific *service periods (SPs)*, in which a PS STA is capable of receiving downlink data traffic. However, each SP is finally terminated *only* by the AP, thus a STA is not able to end such a period on its own.

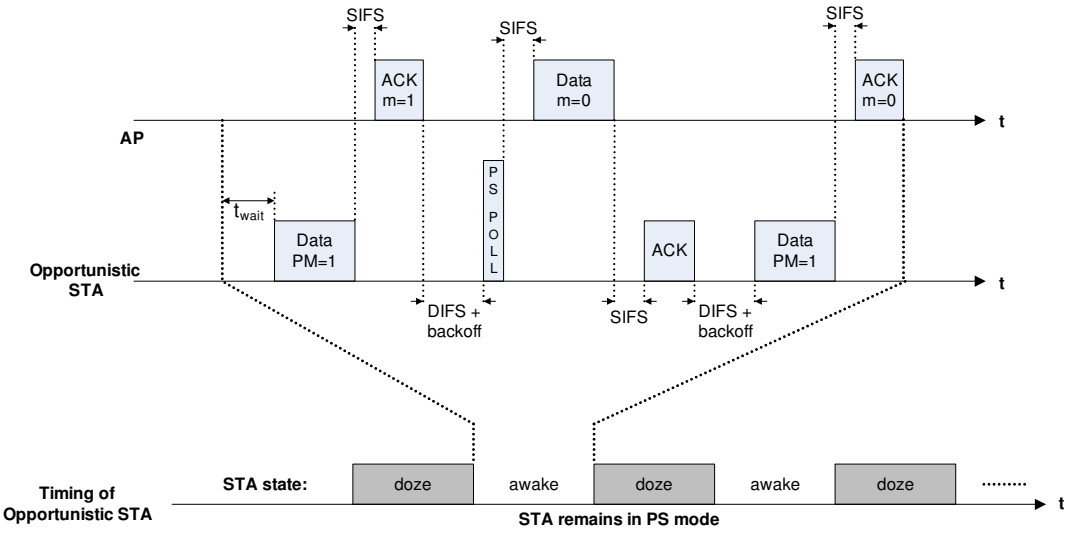


Figure 5.3: WLAN PS signaling for favoring the transport of user data

the AP. As a result, we rely the baseline approach for the opportunistic scheme on two standard compliant steps which enable the STA to keep full control. In a nutshell, these two basic power save features have been already discussed before in Sec. 2.8.2 and 2.8.4: the initiation as well as the termination of the 802.11 power management. Fig. 5.2 shows how we utilize both on small time scales. With the initiation procedure for the power save, we put a WLAN STA quickly into PS mode by setting the PM bit in an uplink frame to one. In contrast, with the termination of the PS mode, we change the PS STA from PS back to active mode (AM). Note that this can be done at any time by signaling STA's mode change to the AP by transmitting an uplink frame with the PM bit set to zero. An opportunistic interval is intended to last only a few to some tens of milliseconds. This short period may lead frequently to the case that our STA does not have any pending uplink data frame transmission for the signaling of the PS mode, thus we simply stick to null-data frames. Again, we point out that the usage of null-data frames for a power-save signaling was widely deployed before, as discussed in Sec. 5.2.3.

5.3.2 Prioritizing the Transport of User Data

To prioritize the transport of VoIP data, we are relaxing the fixed schedule shown in Fig. 5.1. For this, we always complete the transport of all pending data in up- and downlink direction. Afterwards, we immediately switch to the handover preparation phase, which lasts until the beginning of the following opportunistic interval. As discussed in Sec. 5.2.2, in our selected application case for this design option, we passively scan a second WLAN channel during the handover preparation slices.

It may happen, however, that the duration for the data exchange actually exceeds its time slice. To be precise, two issues regarding the timing can appear. First, it

may happen that the residual time slice of an opportunistic interval for a handover preparation is too small to enable the reception of an 802.11 (beacon) frame on the secondary communication channel. In this case, by design, we extend the handover preparation until the beginning of the second subsequent opportunistic interval. Second, the data exchange duration may even exceed one or multiple opportunistic intervals. Then, by design, the handover preparation lasts until the beginning of the first following, subsequent opportunistic interval.

Further, following the prioritization of data traffic on the primary communication channel in a strict way has a tricky implication also on the 802.11 power save mechanisms. There, without the reception of a beacon frame with the TIM (compare Sec. 2.8.3), a STA usually does not know whether one or even more data frames are buffered at the AP awaiting their transmissions in the downlink. Thus, we are required to get a dedicated information whether actually a frame for our STA is waiting at the AP. Note that such information is only obtainable for STAs in *PS mode* via More Data flags (plain 802.11) and "AP PS Buffer State" information (802.11e [6]).

In a nutshell, we stick to the basic frame exchange used for the first design option, but introduce an important difference as shown in Fig. 5.3: To enable a prioritized transmission of VoIP data frames, our opportunistic STA stays *in PS mode all the time* just alternating the STA states periodically between doze and awake. Thus, the null or uplink VoIP data frames have different tasks compared to the first design option. The opportunistic STA sends out the first null-data frame to check at the beginning of each active phase whether downlink data is buffered at the AP in the power save buffer. For a signaling, we utilize an 802.11e option which enables the AP to set the More Data bit *'m'* also in the ACK that acknowledges the reception of the first null-data frame. If the More Data bit is set in an ACK, a standard-compliant STA has to interpret this as if the TIM in a beacon frame indicates the presence of downlink data [6]. Thus, our STA starts to PS poll the AP afterwards, triggering a pending downlink transmission. Note that the More Data bit in this downlink frame also indicates whether other frames are awaiting their transmission in the downlink to our STA. If so, our STA keeps PS polling the AP until all pending data frames have been transmitted. Note that the AP in response to a PS poll either directly replies with a data frame or just acknowledges the poll and delays the actual data transmission. Thus in our approach, the second null-data frame on the one hand checks whether another frame arrived at the AP subsequently. On the other hand, it determines whether the acknowledgment for the last downlink data frame has been received by the AP properly. With this we avoid cases of lost ACKs such that the AP keeps retransmitting the last data frame, while our opportunistic STA is already back in doze mode.

5.4 Methodology for the Performance Evaluation

Our performance evaluation for the opportunistic approach covers the holistic scheme together with the two selected application cases, i.e., the opportunistic scanning as well as the opportunistic preparation of WLAN to WiMAX handovers.

For the holistic scheme, we first conducted an analysis of the timing for *idle channels* given in Sec. 5.5. There, we derived the minimum signaling duration resulting out of the WLAN power save mechanism. On this basis, we analyzed how long it takes to find an existing AP on a given channel, without the presence of any data traffic. Thus, for this analysis, at a first glance, we did not make any assumptions regarding the specific design options discussed in Sec. 5.3 and relied only on the null-data signaling. Then, in a second step, we extended the basic analysis taking up- and downlink data transmissions into account using it also for the signaling of opportunistic scanning. For this, we derived an upper bound for considering the duration it takes to find a given AP, again for idle channels.

As the preliminary analysis considered idle and error-free channels, we further conducted a simulative performance evaluation of the opportunistic scanning presented in Sec. 5.6. There, we analyzed opportunistic scanning in the presence of background load stemming from other WLAN STAs. We evaluated our scheme regarding its scanning duration, the QoS of the on-going VoIP call, and its signaling overhead in the presence of other WLAN STAs, both on the primary and the secondary channel.

Finally, we conducted an analysis of our second application case, the preparation of a handover from WLAN to WiMAX, presented in Sec. 5.7. There, we started by analyzing the timing requirements for both, WLAN and WiMAX. Besides the VoIP transmissions and the power-save signaling, we assumed an idle WLAN channel. Considering in addition an empty WiMAX cell not transporting other traffic streams, we showed that our approach is theoretically feasible over a broad range of WiMAX parameters. Afterwards, we derived thresholds regarding the maximum load level in WiMAX, under which the timing of our opportunistic handover preparation scheme can still be supported.

5.5 Holistic Scheme and Opportunistic Scanning: Analysis for Idle Channels

In this section, we aim at classifying the theoretical performance limits of the opportunistic approach for idle channels both in the first and the second hotspot. Apart from the beacon transmissions and uplink communication between the opportunistic STA with its associated AP, the primary communication channel is assumed to be idle at first. On the secondary channel, we did not consider on-going communication—besides the beacon frames of the present AP. On both channels, we assumed that the selected MCSs for transmissions enable an error-free reception of the layer 2 frames on receiver side. In particular, we intended to answer the following questions:

- How large is the minimum interruption duration for the holistic scheme, including the signaling means described in the design section, and neglecting any downlink data transmissions?

- How long does it take to find an existing AP at a given probability on a specific, idle WLAN channel?

Table 5.1: Size of MPDUs in bytes

G.711 VoIP	null-data	ACK	PS poll	beacon
236	36	14	20	73

Regarding the first question, we analyze the smallest possible duration that is required to initiate and immediately interrupt the preparation of handover steps including all the related signaling. To keep the timing under the control of the STA, in both flavors of our holistic design, we have an uplink transmission before and after a STA starts to conduct the handover preparation. Again, we shall emphasize that the intention for these frames is different in the specific flavors (cf. Sec. 5.3). Nevertheless, this basic signaling duration is identical for both and as such, the analysis is valid for the holistic scheme. From the practical perspective this analysis gives us the smallest possible *black-out duration* that the STA perceives on its actual communication channel thus being not available for any transmissions. This black-out duration is of great interest as it may have a strong impact on an ongoing traffic stream: if the black-out duration is much larger than the service interval of the transported user data, the opportunistic approach may have a negative impact on the end-to-end QoS of the affected traffic stream. Thus, in other words, the minimum signaling duration identifies the magnitude of the smallest service interval of user data being still supportable.

For the second question, we consider a discovery of our second WLAN hotspot on a given channel. The key issue here is to obtain an understanding how often we have to switch back and forth with a selected, fixed time interval until we are able to find a specific AP on a given channel and how long this process takes in time. Although we assume a strong periodicity in the timing, we do not make any assumptions at this stage whether the on-going communication pattern will be prioritized over the discovery or vice versa. We just tackle the effects resulting out of the periodic timing. These results identify the amount of time after which the STA should have found the AP on the given channel. This determines the magnitude of time at which a beacon should have been received, or in other words, after which duration a STA should rather decide to change the channel for the neighbor discovery. The analysis regarding the second question assesses this theoretical limit in the scenario and under the assumptions given below.

Finally, we assess the second question for our application case of opportunistic scanning by considering additionally the presence of downlink data transmissions. Under the assumptions discussed above, we consider the specific power-save signaling of the selected flavor and derive an upper bound regarding the time to scan for a beacon on a selected WLAN channel.

5.5.1 Performance Metrics

In line with the research questions regarding the basic timing principles, we focus on the following three metrics for our analysis: minimum signaling duration, beacon reception probability, and beacon reception duration. The *signaling duration* defines the

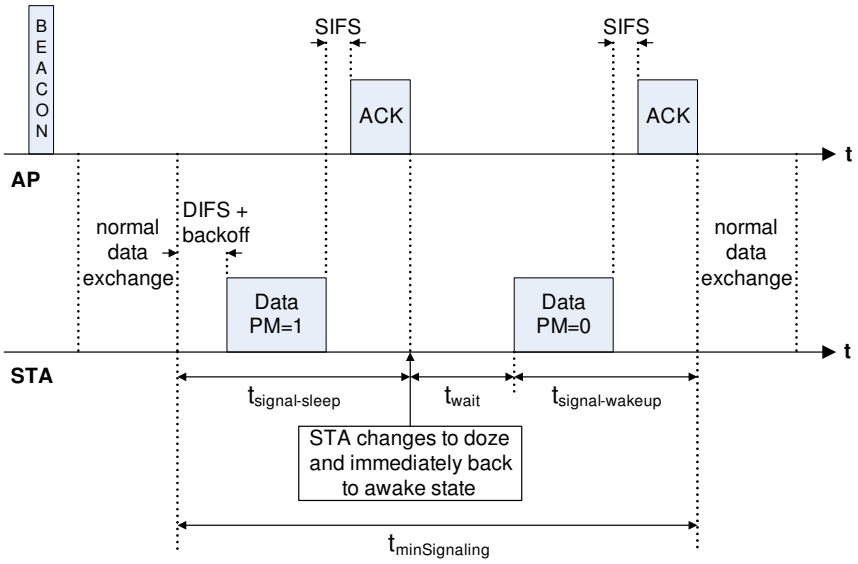


Figure 5.4: Signaling sequence for minimum duration

time span for the initiation and the immediate interruption of the handover preparation steps. It quantifies the service interruption imposed on the application due to the opportunistic approach. The *beacon reception probability* quantifies the number of scan attempts required to successfully receive a beacon at a given probability, whereas the *beacon reception duration* gives the time span of all scan attempts in order to receive a beacon of a single AP on one specific channel.

5.5.2 Minimum Signaling Duration

This section analyses the minimum signaling duration that is required for the initiation and the immediate interruption of the opportunistic preparation steps afterwards. Figure 5.4 illustrates the signaling sequence involved in going from awake into doze and immediately back into awake state. As we are interested in the minimum signaling duration, we do not spend any time in the doze state, thus we are actually not conducting any opportunistic preparative steps at all. In other words, this quantifies the smallest possible duration to switch back and forth between channels for both flavors of our holistic scheme. In order to hold a specific QoS constraint for a certain type of traffic, i.e., to avoid a backlog of data frames on STA side, we consider the minimum signaling duration as a lower bound for the inter-arrival time of application data packets at MAC level. The minimum signaling duration ($t_{\text{minSignaling}}$) is given by

$$t_{\text{minSignaling}} = t_{\text{signal-sleep}} + t_{\text{wait}} + t_{\text{signal-wakeup}}, \tag{5.1}$$

whereby the individual components are determined as follows:

$$
\begin{aligned}
t_{\text{signal-sleep}} &= t_{\text{DIFS}} + \text{rand}_{\text{uniform}}(0, \text{cw}) \cdot t_{\text{slot}} + t_{\text{DATA-UL}} + t_{\text{SIFS}} + t_{\text{ACK}}, \\
t_{\text{wait}} &= \begin{cases} t_{\text{probeD}}, & \text{if channel is idle,} \\ t_{\text{busy}} + t_{\text{DIFS}} + & \\ \quad + t_{\text{rand(0,cw)}} \cdot t_{\text{slot}}, & \text{if channel is busy,} \end{cases} \\
t_{\text{signal-wakeup}} &= t_{\text{DATA-UL}} + t_{\text{SIFS}} + t_{\text{ACK}}.
\end{aligned}
$$

Assuming an idle channel, Equation (5.1) can be simplified to

$$
\begin{aligned}
t_{\text{minSignaling}} &= t_{\text{DIFS}} + 2 \cdot t_{\text{SIFS}} + 2 \cdot t_{\text{DATA-UL}} + \\
&\quad + 2 \cdot t_{\text{ACK}} + t_{\text{probeD}}.
\end{aligned} \tag{5.2}
$$

Apart from PHY specific parameters (t_{DIFS} and t_{SIFS}), $t_{\text{minSignaling}}$ depends on the employed MCS for the data and the acknowledgment frames [6]. Further, t_{probeD} is the probe delay ensuring that the STA does not harm any on-going communication. Work on practical implementations reported WLAN STAs to be capable of starting transmissions after 400 to 800 μs on a channel, when devices changed from doze to awake state [133, 134]. Thus, we decided to select a duration of 1024 μs, being rather conservative in our choice.

Figure 5.5 shows the minimum signaling duration for different 802.11 PHYs. The results account for two situations: In the first, we assume that all signaling is conducted by means of uplink (UL) null-data frames. By contrast, in the second, we piggyback the signaling completely in uplink VoIP data frames. For this we assume to have typical VoIP traffic with short inter-packet times resulting from the G.711 codec [120] with 20 ms packetization not considering silence suppression (leading to 160 Bytes voice data for each 20 ms). Table 5.1 gives the size of the MAC protocol data units (MPDUs) including packet payload and MAC header used for analysis (and subsequent simulations).

Obviously, the smallest achievable interruption of 1.23 ms occurs for the lowest packet size (null-data frame) at the highest data rate for an 802.11g PHY. But also with the most robust MCS scheme of 802.11b, i.e., 1 Mbps, just a 2.7 ms-long interruption appears. Including the signaling piggybacked in uplink data, the minimum duration further increases as a result of the data payload. For the considered VoIP traffic with the G.711 codec and 20 ms packetization together with the most robust MCS scheme of an 802.11b PHY, the service interruption lies in the order of 5.9 ms. These analytical results show that our approach may not noticeably affect the VoIP application. The smallest service interruption is significantly lower than the packetization time—both with null-data frames as well as with piggybacked signaling within the VoIP packets.

5.5.3 Required Scan Duration

In the following, we analyze how long it takes to find an existing AP at a given probability on a specific, idle channel. In order to detect a neighboring AP during the nth + 1 opportunistic scan attempt, the beginning of the scanning time (scanning start, t_{SS}) has

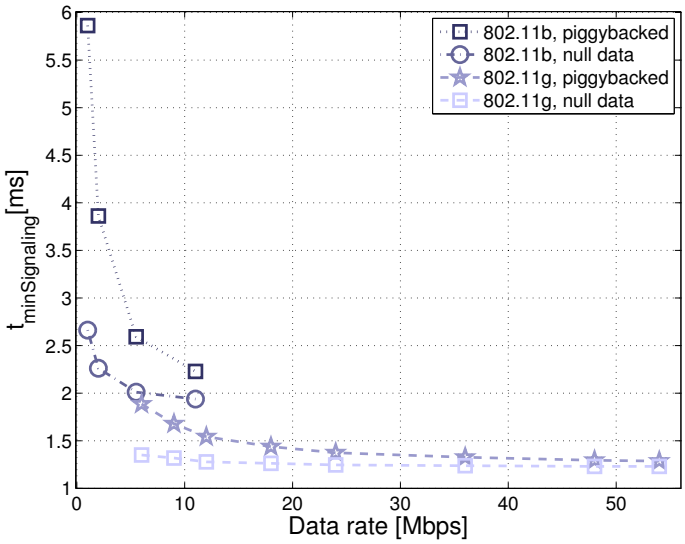


Figure 5.5: Minimum signaling duration for 802.11b HR/DSSS and 802.11g ERP-OFDM PHYs, UL-signaling is either realized by null-data frames or piggybacked in VoIP frames

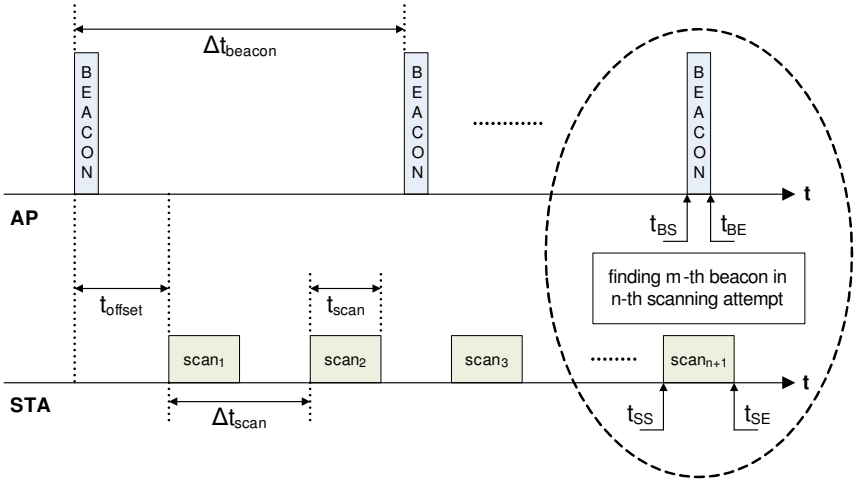


Figure 5.6: Calculation of the number of scan attempts (signaling not shown)

to lie before the actual beginning of the beacon frame (beacon start point, t_{BS}), while the end of the scan attempt (scan end point, t_{SE}) has to occur after the end of the beacon frame (beacon end t_{BE}) (c.f. Fig. 5.6):

$$t_{\mathrm{SS}} \leq t_{\mathrm{BS}} \ \wedge \ t_{\mathrm{BE}} \leq t_{\mathrm{SE}}. \tag{5.3}$$

Therein, we substitute t_{SS}, t_{BS}, t_{BE}, and t_{SE} as follows

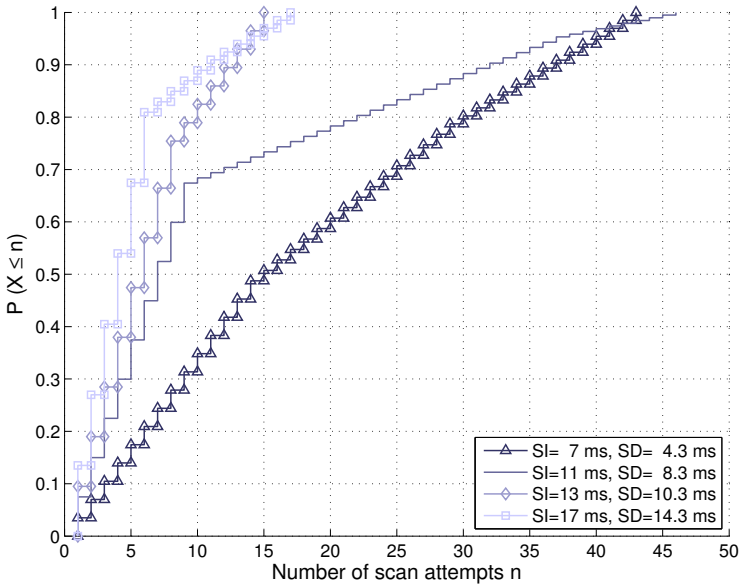
$$\begin{aligned} t_{\mathrm{SS}} &= t_{\mathrm{offset}} + n_{\mathrm{scan}} \cdot \Delta t_{\mathrm{scan}}, \\ t_{\mathrm{BS}} &= n_{\mathrm{beacon}} \cdot \Delta t_{\mathrm{beacon}}, \\ t_{\mathrm{BE}} &= t_{\mathrm{BS}} + t_{\mathrm{beacon}}, \\ t_{\mathrm{SE}} &= t_{\mathrm{SS}} + t_{\mathrm{scan}}, \end{aligned}$$

where t_{offset} is a random variable uniformly distributed over $[0, \Delta t_{\mathrm{beacon}})$, $\Delta t_{\mathrm{beacon}}$ the beacon interval, t_{beacon} the duration of a beacon frame, and Δt_{scan} the opportunistic interval being denoted as scan interval (SI) in the following. Further, t_{scan} is the duration for the handover preparation, which we denote as the (effective) scan duration (SD). Again, it is the duration remaining after the involved signaling is deducted from the time span given by Δt_{scan}. Equation 5.3 can accordingly be rewritten into

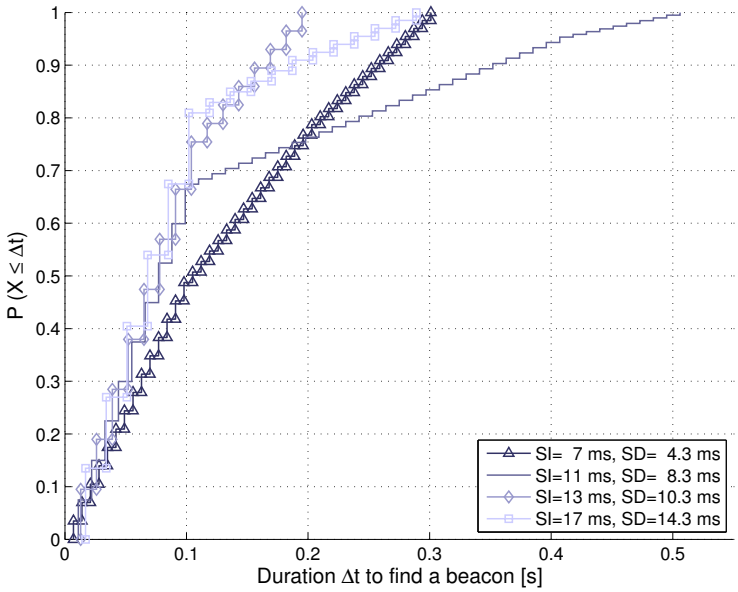
$$\begin{aligned} & \tfrac{n_{\mathrm{beacon}} \cdot \Delta t_{\mathrm{beacon}} - t_{\mathrm{offset}}}{\Delta t_{\mathrm{scan}}} - \left(\tfrac{t_{\mathrm{scan}} - t_{\mathrm{beacon}}}{\Delta t_{\mathrm{scan}}} \right) \leq \ n_{\mathrm{scan}} \\ \wedge \qquad & n_{\mathrm{scan}} \ \leq \ \tfrac{n_{\mathrm{beacon}} \cdot \Delta t_{\mathrm{beacon}} - t_{\mathrm{offset}}}{\Delta t_{\mathrm{scan}}}, \end{aligned} \tag{5.4}$$

which gives the condition that beacon number n_{beacon} is successfully received within scan attempt n_{scan}. We analyze the latter equation numerically in an iterative fashion for different values of t_{offset}. Due to the stochastic nature of t_{offset}, we obtain the probability functions of detecting a beacon at a given scan attempt / after a given time (c.f. Fig. 5.7 and Fig. 5.8). Obviously, t_{offset} and $\Delta t_{\mathrm{beacon}}$ may not have a common divider to guarantee beacon detection. As we assume that a WLAN operator will employ common values with multiples of 10 ms for the target beacon transmission time (e.g., 100 ms) we choose prime numbers for Δt_{scan}. By this, we avoid recurrent cases in which beacon transmissions always appear while a STA is not in the scan phase.

In a first step, we analyze the case if only uplink null-data frames are used and no other traffic is present. This happens frequently for SIs being much smaller than the packet intergeneration time of a transported traffic pattern. The results are given in Fig. 5.7. We observe that longer SIs yield to better results regarding the number of required scan attempts as shown in Fig. 5.7a. Interestingly, the effect is less noticeable if one considers the time required to find a beacon in Fig. 5.7b, as compared to the number of scan attempts. The discovery of an AP within two beacon intervals is possible in 75 percent of the cases even for the smallest SI of 7 ms. This is only twice the time needed as compared to traditional passive scanning resulting into long service interruptions. From the results, we can further identify the impact of unsuitable SIs such as 11 ms, which resolve the periodicity with the beacon intervals slowly. In such cases, although resulting in a high duration, this SI can accomplish a successful discovery within five beacon intervals.

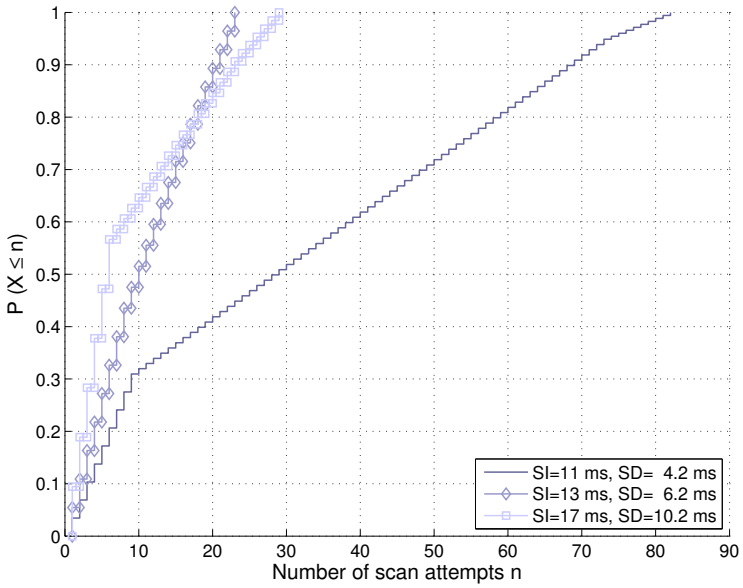


(a) Empirical CDF of the scan attempts

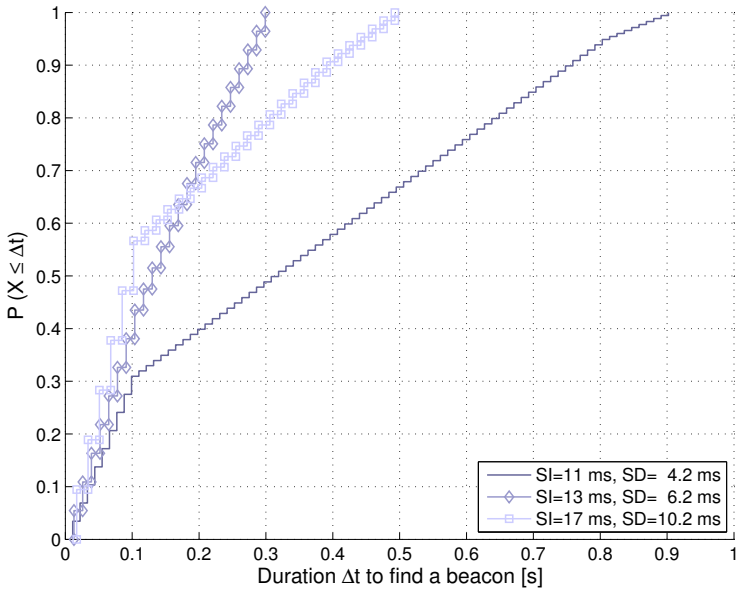


(b) Empirical CDF of the scan duration

Figure 5.7: Probability of receiving a beacon; the AP applies a 100 ms beacon interval; scan intervals (SIs) at the STA range from 7 to 17 ms with given scan durations (SDs); signaling with null-data frames, all transmissions with 802.11b HR/DSSS@1 Mbps.



(a) Empirical CDF of the scan attempts



(b) Empirical CDF of the scan duration

Figure 5.8: Probability of receiving a beacon; the AP applies a 100 ms beacon interval; scan intervals (SIs) at the STA range from 11 to 17 ms with given scan durations (SDs); in each SI, an up- and a downlink VoIP packet is pending, all transmissions with 802.11b HR/DSSS@1 Mbps.

5.5.4 Opportunistic Scanning: Upper Bound of Scan Attempts

Next, we study the mix of null-data signaling and G.711 VoIP traffic in up- and downlink with a 20 ms packetization. For an upper bound analysis regarding the overall time to find a beacon, we assume to have in each SI one VoIP data packet pending in up- as well as downlink. We apply all means of our second design flavor discussed in Sec. 5.3.2 including the PS polling of a downlink data frame. Then, the complete data exchange duration lasts for 6.8 ms. Note that for this upper bound with a pending up- and downlink transmission in each SI, the 7 ms SI is not practical as it leaves about 0.2 ms for scanning other channels. Fig. 5.8 shows the number of required scan attempts and the time to find a beacon for the remaining SIs. Due to the VoIP transmissions, the SD reduces by about 4 ms in each SI compared to the case with null-data frames. As a result, also the number of scan attempts and the overall duration of the opportunistic scanning increases. Nevertheless, a beacon is found for an SI of 13 ms within 300 ms, while the SI of 11 ms results in 900 ms (compared to 200 and 500 ms for the case with null-data frames). Again, we emphasize that these results serve as an upper bound as VoIP traffic with a packetization of 20 ms rarely results in pending up- and downlink traffic in each of these small SIs. Note that this analysis applies only, if all three aspects, namely the beacon to be scanned as well as a transport of VoIP frames in up- and downlink, together fall in the *same* SI.

5.5.5 Summary of the Timing Analysis for Idle Channels

In a first step, we presented a timing analysis of the novel opportunistic approach employing 802.11 power management as a underlying signaling protocol. We showed that depending on the employed MCS scheme, service interruption may be reduced to values in between 2 and 5.9 ms determined for 802.11b with null-data frames at 11 Mbps and the piggybacked signaling at 1 Mbps. This makes the opportunistic approach attractive for real-time communication with tight QoS constraints while being entirely compliant with the 802.11 standard. The basic analysis, which aims at a discovery of a second WLAN hotspot while supporting an on-going VoIP call, shows for an idle channel and pure null-data signaling, a beacon reception is doable in less than 200 ms. Next, considering the opportunistic scanning, a mix of null-data and VoIP traffic increases the time to find a beacon at most up to 300 ms. Thus, opportunistic scanning is only twice to three times as long as compared to the traditional passive scanning of an AP with a 100 ms beacon interval. Overall, these results give us a solid basis for a further evaluation of the opportunistic approach. By means of results from simulations, we reveal the effects of channel load caused by other stations and evaluate the costs of our approach in terms of signaling overhead in the following section.

5.6 Opportunistic Scanning: Performance Evaluation with Background Load

The preliminary analysis of the basic timing principles for opportunistic scanning presented in the previous section focussed on idle channel conditions, thus considering beacons arriving timely at TBTTs. Further, the actual costs of the opportunistic approach in terms of signaling overhead have not been discussed so far. As a result, we conduct a performance evaluation of these aspects by means of simulations. Thereby, we aim to quantify the behavior of the scheme applied to an 802.11 WLAN system used for real-time VoIP communication. Accordingly, we apply the second design option (cf. Sec. 5.3.2), with the application case of opportunistic scanning, as we prioritize the transport of the VoIP packets over the neighbor discovery in order to uphold the corresponding QoS limits. For this, we follow a methodology consisting of three main aspects: First, we consider an idle channel (no other STAs competing for channel access). This isolates the effect of the scanning scheme itself on the real-time traffic. Second, we reveal how background traffic further influences the performance and sturdiness of our novel scanning scheme. Third, we quantify the overhead associated with opportunistic scanning.

The evaluation of the *performance for idle channel conditions* acts as a baseline case showing how opportunistic scanning itself influences the inter-arrival time of user data packets. Also, we show how long our approach needs to discover one existing AP at a given probability. Note that on idle channels, the passive discovery of an AP takes the longest time as only beacon frames can be 'sniffed'. In a second step, we aim at studying the *influence of background traffic* showing how it affects opportunistic scanning: a potential busy channel will not only delay the transmission of user data, but also hinder the (piggybacked) signaling information, and even procrastinate the transmission of beacons being scheduled at fixed time-intervals. Further, this performance evaluation *quantifies the protocol overhead* in terms of additional signaling information imposed by opportunistic scanning.

All the former performance aspects are evaluated by simulations considering all relevant protocol aspects of 802.11 [6] such as CCA functionality, random backoff, delayed beacons due to busy media, and transmit-receive-turnaround time of the RF front-end.

The contributions in this section can be summarized as:

- assessing implications of the 802.11-architecture and protocol such as delayed beacons and clear channel assessment accounting for random backoff due to a busy media,
- comparing results for an idle communication channel to the analytically derived performance limits of Sec. 5.5,
- evaluating the influence of background traffic on the performance of opportunistic scanning, and
- quantifying the costs of opportunistic scanning including a detailed discussion of the introduced protocol overhead.

5.6.1 Metrics

The following metrics are used throughout the performance evaluation:

- *Beacon reception probability* quantifies the number of scan attempts / time required to successfully receive a beacon at a given probability.
- *Inter-arrival time (IAT)* of user data is the time between consecutively received data packets transmitted in the downlink.
- *Packet loss* accounts for the percentage of lost (user data) packets in the downlink direction.
- *Average scan duration* identifies the average time spent on a single scan attempt on the neighboring AP's channel.
- *Average data exchange duration* measures the time required to transmit and receive all buffered (user data) packets, null-data and PS poll frames for each SI.
- *Null-data frame rate* quantifies the protocol overhead as null-data frames are used to check whether downlink data is buffered at AP, if no pending user data is awaiting its transmission.
- *PS poll frame rate* gives the protocol overhead as the STA polls buffered downlink data when being in PS mode.

As some of our metrics base on averages, we conduct simulations with *independent replications* [153]. For this, we repeated the simulation runs hundred times. Further, we calculated the 90 percent confidence level for these metrics.

5.6.2 Simulation Scenario

Modeling our reference scenario, our simulation scenario consists of two adjacent APs operating on non-interfering channels, having an overlapping coverage area, and a connection to the Internet (c.f. Fig. 5.9). Both APs transmit beacons to announce their BSS. Again, both APs do not coordinate or even synchronize the process of sending the beacon frames. Beacon transmissions are scheduled at each AP at own regular time intervals as defined by the 802.11 standard [6] but may be delayed if the AP's CCA functionality indicates the medium to be busy. Both APs are connected via Ethernet to a VoIP server; whereby, the wired connection does not impose any throughput constraints.

The STA conducting opportunistic scanning is placed in the overlap of the adjacent cells. It is associated with one of the two APs which is also used to relay an ongoing VoIP call (G.711 codec employing 20 ms packetization without silence suppression). Employing opportunistic scanning for network discovery, the STA senses on the neighbor AP's channel for beacon transmissions.

Background traffic is increased by continuously adding additional STAs each featuring one ongoing VoIP call, which we denote as *background-load STAs (BG STAs)* in the

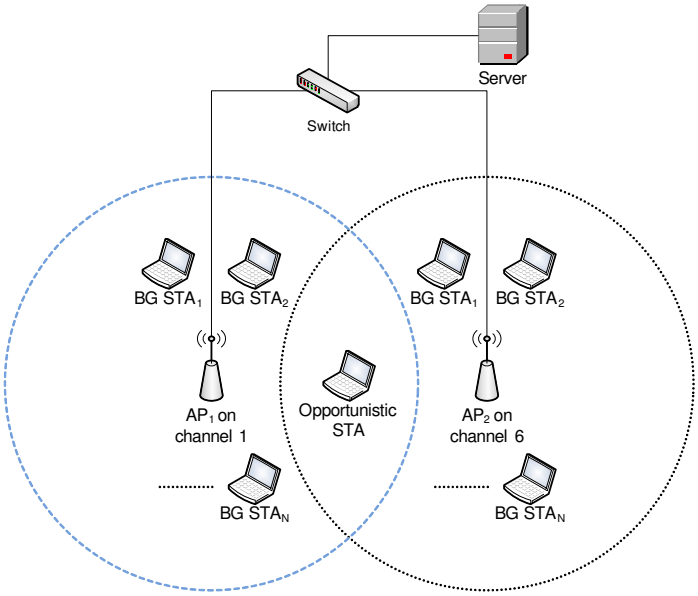


Figure 5.9: Simulation scenario, static STAs with background load

following. In each step, one BG STA is added per AP and is placed within the latter's coverage such that STAs are in communication range of each other. This avoids hidden terminal effects as RTS/CTS is disabled for the considered VoIP service. By this, we clearly isolate the opportunistic scanning scheme to be the cause of any observed effects.

We increase the number of BG STAs until the QoS limits of the VoIP traffic are violated. Following the results of Cole and Rosenbluth [154] regarding the impact of packet delay and losses on the VoIP quality, we consider QoS limits in terms of a maximum end-to-end delay of 200 ms and a maximum packet loss of 4 percent.

5.6.3 Simulation Model

For our simulations, we used the OPNET Modeler Wireless Release 14PL2 [155]. Thereby, the basic 802.11 model of the simulation library was extended by the power saving features discussed in Sec. 5.3.2.

Let us briefly highlight the changes on MAC level to the basic 802.11 model of OPNET. We included power save states and modes in the 802.11 state machine. Further, we added the PM bit in MAC headers, and introduced the formats of null-data and PS poll frames. Lastly, we adapted the queuing for the AP. In the original 802.11 model, the AP has one queue to store the downlink traffic of all its associated STAs in a FIFO fashion. For a support of the power save operation, we introduced a separate queue with a size of 20 data packets for each associated STA in PS mode. If the AP receives a PS poll of one of these STAs, it responds with a data packet from the corresponding queue. If the queue is not empty, we set the More Data bit in the data frame.

On PHY level, we assumed a switching time of 25 μs [156] to change from one WLAN channel to another. Further, we used the 802.11b amendment and applied the HR/DSSS PHY with long preambles as given in Sec. 2.5.2. For the selected PHY, we considered 802.11b data rates of 1 and 11 Mbps to enable a comparison with previous work, specifically regarding the VoIP capacity of our setup.

5.6.4 Simulative Results for Idle Channel Conditions

During the numerical analysis in Sec. 5.5 we argued that prime numbers shall be chosen for the SI in order to guarantee a successful reception of a beacon after a finite time span. Further, our analysis has shown that SIs of 7 ms may leave only a negligible scan duration for 802.11b PHYs, while 11 ms SIs are only slowly resolving periodicity issues with 100 ms beacon intervals. Accordingly, we show results for prime SIs in between 13 ms and 51 ms. The smallest value guarantees a service rate less than the common VoIP packetization rate of 20 ms and the largest is in the order close to the 50 ms service interruption acceptable for voice communication. To enable a direct comparison with results from the analysis, we first applied the 802.11 HR/DSSS PHY with 1 Mbps, only.

Impact on Inter-Arrival Time of Data Packets

We analyzed how opportunistic scanning impacts the IAT of received user data packets at the STA in the downlink. Again, opportunistic scanning coarsely divides the time for each STA independently into a scanning period and a data exchange period. As this pattern is periodically repeated at a fixed rate, one intuitively expects VoIP packets to arrive at the STA at multiples of the SI. Figure 5.10 shows the cumulative distribution function (CDF) of the IATs for a SI of 13 ms. For the 20 ms packetization rate of the VoIP application, the two modes of the IAT's distribution lie expectingly at the two multiples closest to the SI, namely 13 ms and 26 ms. Also, we observe that slightly more packets are transmitted within every second SI. The reason for this actually lies in the relationship of the chosen SI and the packetization of the VoIP traffic: The 26 ms multiple of the SI is simply closer to the expected IAT of 20 ms as the 13 ms multiple. Choosing a different SI of 17 ms would result in slightly more transmissions to occur at 17 ms and 34 ms.

Comparison of Analysis and Simulation Results

For idle channels, Figure 5.11 illustrates the probability to successfully detect the neighbor AP's beacon after a given time span. For a comparison, we also plotted the analytical results from Sec. 5.5.3 with the null-data transmissions. The effects of random backoff, switching times of the RF front-end, the PS polling, and the data transmissions impose only a marginal difference. Thus, for idle channel conditions, we suggest to predict the distribution of the scanning duration, depending on selected beacon and scanning intervals, simply on the basis of the analytical form given in Sec. 5.5.3.

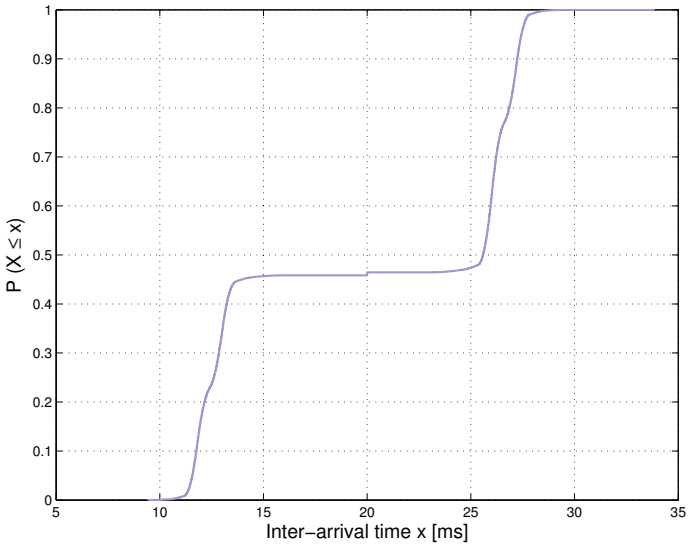


Figure 5.10: Influence of opportunistic scanning on the IAT of downlink VoIP transmissions, shown as CDF over the IAT, for an SI of 13 ms, 802.11b HR/DSSS PHY with 1 Mbps for both, analysis and simulations.

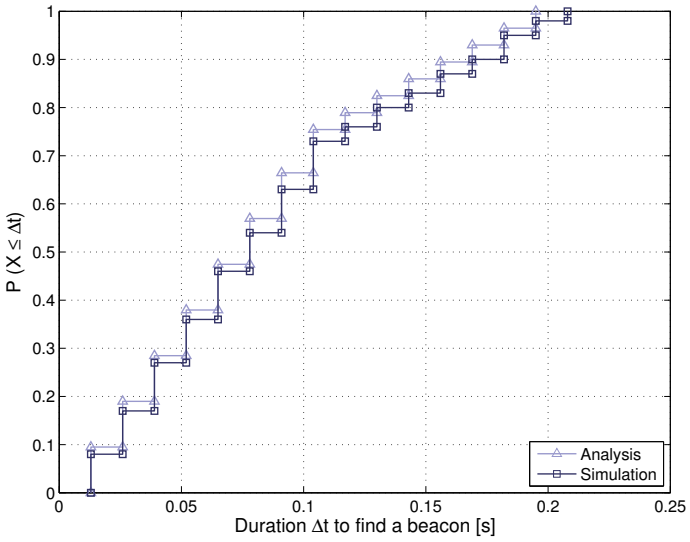


Figure 5.11: Time to find a beacon: comparison of CDFs from analysis and simulations, for an SI of 13 ms.

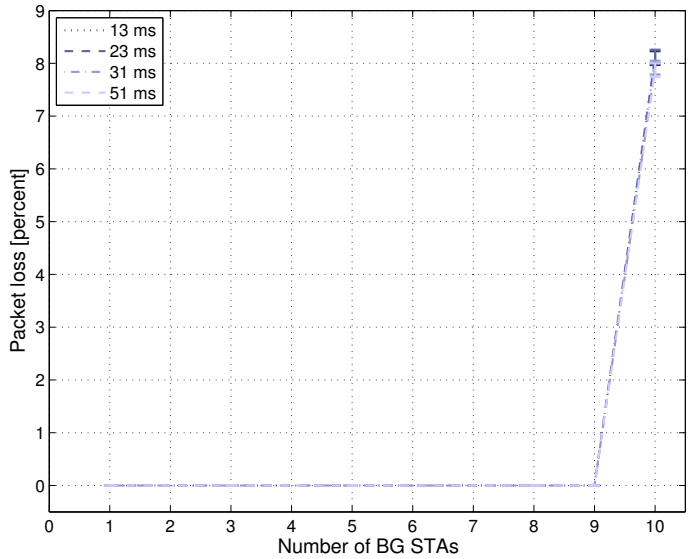


Figure 5.12: Packet loss probability for downlink VoIP frames

5.6.5 Influence of Background Traffic

Opportunistic scanning is vulnerable to background traffic as an increased network load within a BSS may cause delayed beacons. In addition, the start of an opportunistic scan attempt initiated by our signaling prior a change from *awake* to *doze* state may be delayed due to a busy medium. As our scanning scheme primarily aims at providing a data exchange opportunity at every SI to enable continuous communication at application level, increasing the signaling duration automatically reduces the effective time available for scanning another channel.

Maximum Number of VoIP STAs

In a first step, we determined the maximum number of VoIP STAs that can be simultaneously handled by an AP. Figure 5.12 illustrates the packet loss probability if in addition to a STA applying our opportunistic approach, further VoIP STAs are added to the BSS. We notice no packet loss up to a total of 10 STAs in the system (9 BG STAs and one opportunistic STA). Having 11 ($= 10 + 1$) STAs in the system saturates the WLAN and we notice packet loss due to buffer overflows at MAC level violating the QoS constraints. Considering own work [157] as well as in addition results from other authors [158] regarding the VoIP capacity of an 802.11b WLAN cell, our novel scanning scheme reduces the maximum number of supportable STAs only by one at the gain of guaranteeing QoS at application level and conducting a continuous network discovery.

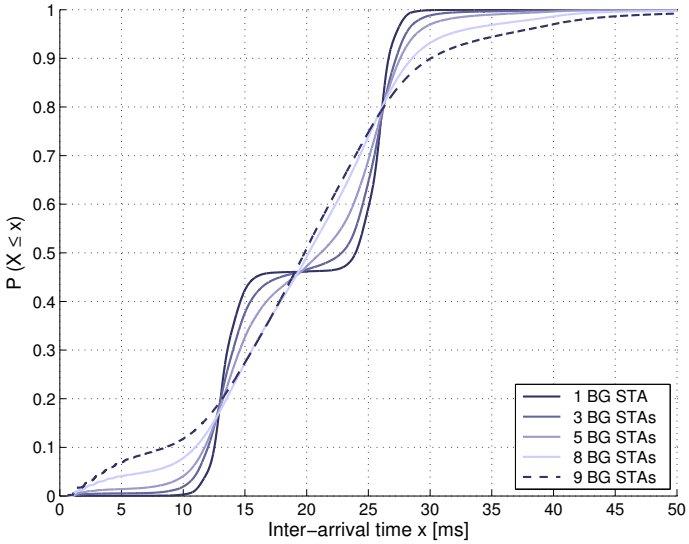


Figure 5.13: Influence of background traffic on the IATs of downlink VoIP transmissions, for a SI of 13 ms, 802.11b HR/DSSS PHY with 11 Mbps data frames.

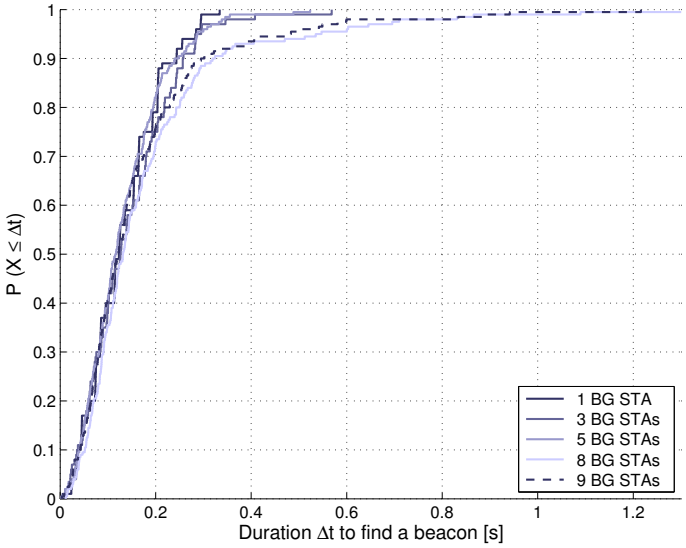


Figure 5.14: Influence of background traffic on total scan time, for a SI of 13 ms, 802.11b HR/DSSS PHY with 11 Mbps for data frames.

Impact on Inter-arrival Time for VoIP Packets

For an increased background load, the channel access delay rises and is more and more dominated by the 802.11 backoff procedure. Hence, the data exchange between the AP and the opportunistic scanning STA is less deterministically bound to multiples of the scanning interval as shown in Fig. 5.13. The distribution of the IAT of VoIP packets changes from a two-modal towards a uni-modal shape the closer the network load comes to the maximum number of VoIP STAs in the system. Further, we notice an increase of the 99-percentile for the inter-arrival time from 28 ms to 45 ms.

Effect on Scanning Duration

The increased background load results in longer periods spent for the exchange of user data per SI. Thus, less time is actually left per scan attempt which in turn reduces the beacon reception probability. Figure 5.14 illustrates that the 99-percentile increases from 300 ms for a single VoIP background flow up to 0.85 to around 1.1 s for the highest achievable system load. For the selected type of background load, this time represents the longest possible duration for the detection of another cell on a given channel while still upholding an ongoing, interruption-free communication on application level. For completeness, Figures 5.15 and 5.16 illustrate the effect of an increased background load both on the effective scan and the data exchange duration for different SIs. It shows that opportunistic scanning is not able to complete the data exchange and the scanning within an SI, especially for small SIs at increasing load levels. Essentially, this is a result from the selected design of the opportunistic scanning. The preference in this approach is given to the transport of the user data, i.e., data exchange durations are extended at the cost of reduced scan durations as discussed in the specific design in Sec. 5.3.2.

5.6.6 Quantification of the Protocol Overhead

In the previous subsection, we already quantified the overhead associated with opportunistic scanning: the maximum number of supportable VoIP STAs within a BSS is reduced by one. In the following, we highlight the cause for this reduction.

If data is pending for uplink transmission, opportunistic scanning does not impose any overhead. Only if the STA does not have data ready to transmit in the uplink, it sends a null-data frame to trigger a check whether downlink data is buffered at AP. Figure 5.17 illustrates the influence of the background load on the null-data frame rate for various SIs. The smallest considered SI of 13 ms results in a significantly higher overhead, because the 20 ms packetization rate of the VoIP flow is larger than the SI. Hence, this results frequently in an empty data queue on STA side, if a trigger for a check of buffered downlink data is pending. For the other SIs, the influence of the background load is almost negligible, as the load-dependent increase of null-data frames remain low.

Although our approach uses entirely standard compliant means, it has one disadvantage: each packet buffered at the AP for downlink transmission has to be requested by the STA using a PS poll frame. Accordingly, we expect a PS poll frame rate of 50 frames

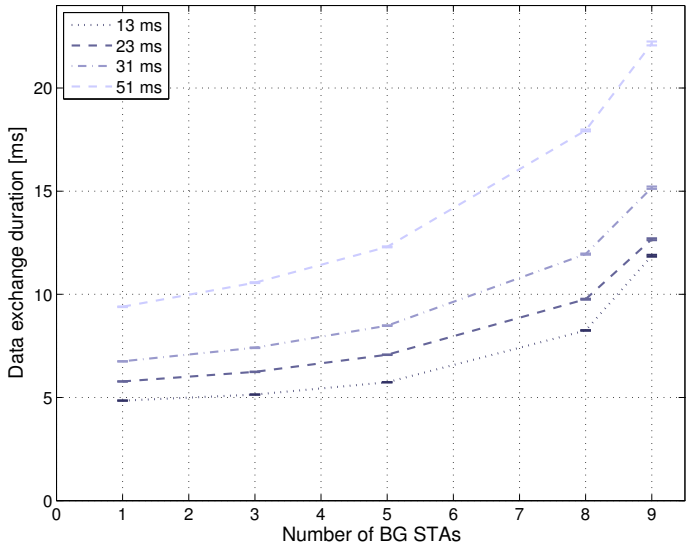


Figure 5.15: Average data exchange duration of an opportunistic STA

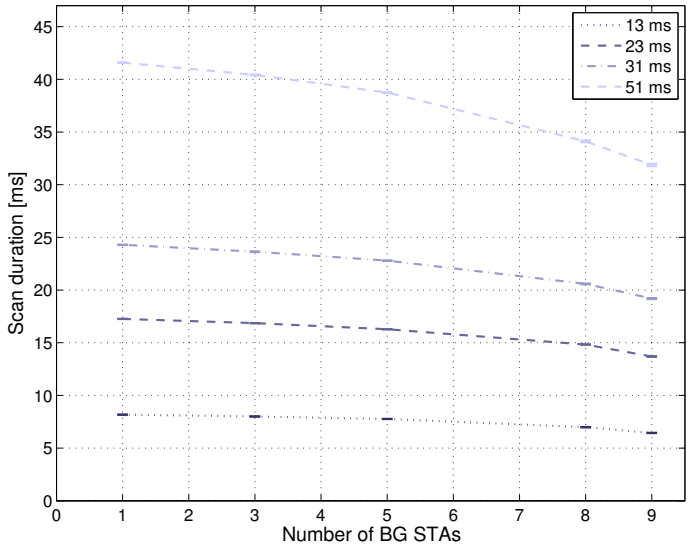


Figure 5.16: Average scan duration of an opportunistic STA

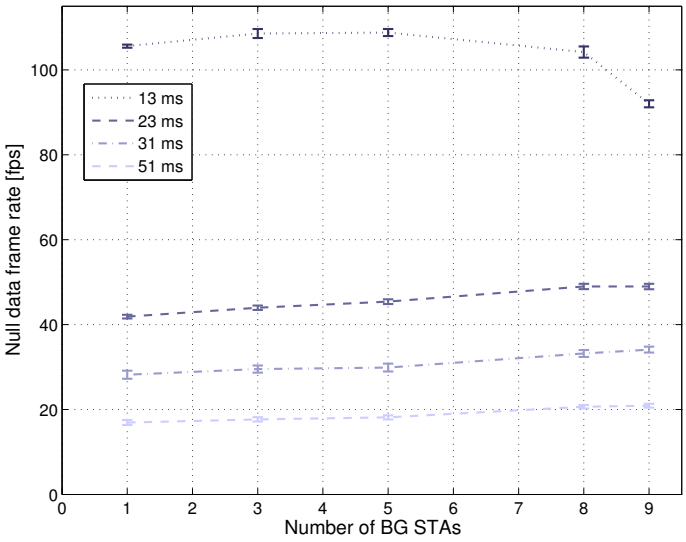


Figure 5.17: Number of sent null-data frames per second

per second, corresponding with the number of VoIP packets in the downlink. As shown in Fig. 5.18, the PS poll frame rate slightly increases with the number of background STAs being a result of an increased collision level, and requiring on average about one additional PS poll transmission per beacon interval of 100 ms.

5.6.7 Summary of Simulation Results

This section presented a comprehensive performance evaluation of the opportunistic scanning using VoIP as an example for a real-time communication. We showed that opportunistic scanning results in a bi-modal distribution of user packet inter-arrival times. Furthermore, the 99-percentile of the observed inter-arrival times lies below 40 ms for high background load levels. Thus, we conclude that the variance of the packet delay introduced by our opportunistic scanning approach may only have a marginal impact on the end-to-end QoS of the considered VoIP stream as this effect is usually filtered out by play-out buffers in the VoIP stack. Featuring a *network discovery* even in the presence of background load, opportunistic scanning is fully capable of supporting real-time applications such as VoIP if a duration of up to 1.1 s is acceptable for the specific use case of the discovery process of an AP on a given channel. Cost-wise, the maximum number of simultaneously supportable users in the WLAN cell is only reduced by one. We conclude that opportunistic scanning is capable to continuously monitor the environment of a STA while upholding QoS constraints for real-time services. Hence, it is suitable as a network discovery scheme for various application scenarios such as periodic background scans.

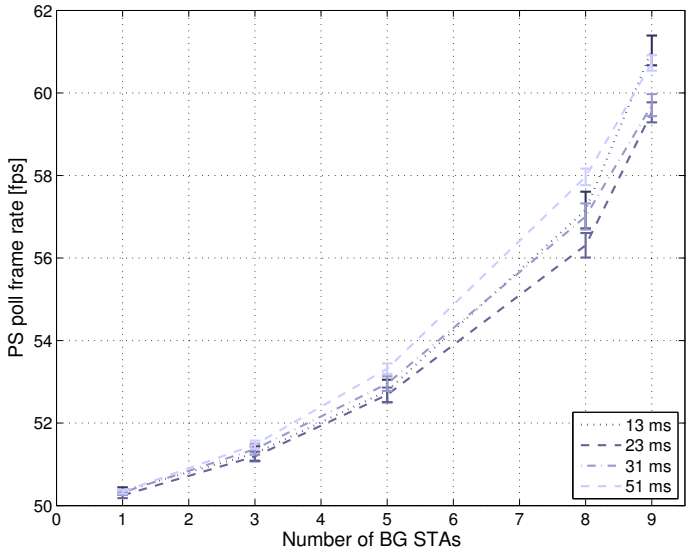


Figure 5.18: PS poll frame rate

5.7 Preparation of WLAN to WiMAX Handovers: Timing and Load Analysis

Let us now consider our application case for handover preparation steps in a heterogeneous environment belonging to the first design category. As such, this section deals with the problem of how to enable a usage of heterogeneous links over a multi-mode NIC while meeting QoS constraints of an on-going transmission—including constraints imposed by real-time voice connections. Thereby, besides the VoIP traffic and the power-save signaling for the opportunistic device, we assume to have an idle channel in the WLAN hotspot. On both, the primary and the secondary communication channel, we further assume that the selected MCSs for transmissions enable an error-free reception of the layer 2 frames on receiver side. We present a solution compatible to the family of 802.11 and 802.16 standards, whereby we have selected the latter as it is well-known for its lengthy network entry [142]. We identified the fundamental limits of the second flavor of the opportunistic approach by considering the timing issues for both WLAN communication and WiMAX network entry. We analyzed the following questions:

1. Under the assumption of an idle WLAN channel, how long is the maximum duration of the data exchange including the power-save signaling?

2. On the basis of the WLAN timing from the first question, can the opportunistic approach fulfill the timing requirements of a mobile WiMAX cell with an empty channel? If so, what would be the theoretical load limits for WiMAX still enabling a support our scheme?

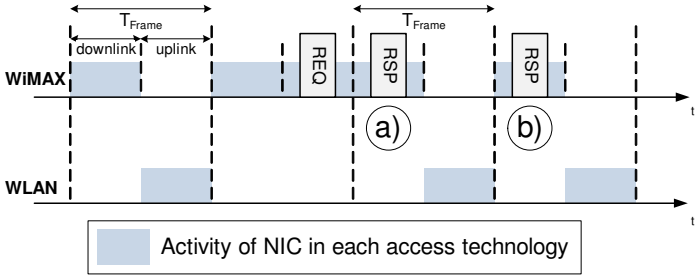


Figure 5.19: WLAN-WiMAX alternation principle

5.7.1 Basic Principle

Following the basic idea of the opportunistic approach, we pause WLAN by means of the power save mode and switch to WiMAX in the gaps (Fig. 5.19). Let us analyze the 'quanta' in which the basic WiMAX entry has to proceed, as given by the IEEE 802.16 standard [143] and summarized in the work of Hollick et al. [142]. The first step consists of a *discovery* of the WiMAX network by scanning for preambles of downlink subframes. Afterwards, the WiMAX device adapts to the strict timing of the WiMAX time frames. This is followed by the steps of *obtaining downlink (DL)/uplink (UL) parameters, initial ranging, capability negotiation, authorization and key exchange, and registration.* Lastly, the *establishment of the IP connectivity and a service flow* finally prepares WiMAX for the VoIP transport [142, 143].

In order to allow for our opportunistic STA on the one hand a fast setup of the second path via WiMAX and on the other hand to be able to adapt to the strict timing pattern of the WiMAX MAC, we give a strict timing priority to WiMAX. For this, we apply the following basic principle. There, the wireless device has to be present during the DL part for *all* WiMAX frames. In case there is no pending action for the UL, the device switches to WLAN and returns back for the start of the next WiMAX frame.

Beside the neighbor discovery, all further steps of the network entry for WiMAX [142] are based on *request (REQ) / response (RSP)* two-way handshakes, where the device issues the REQ and waits for the RSP of the BS. For each of these steps, the wireless device has to spend the complete frame plus the following DL subframe in WiMAX mode (in order to send out REQ and potentially receive RSP, if transmitted by BS immediately in the subsequent DL subframe, Fig. 5.19 case a)). If RSP will be sent later in one of the following frames, it will be received by the device anyway since it always spends the DL-part of the frame within WiMAX (Fig. 5.19 case b).

The timing priority of WiMAX may lead to UL phases, in which the device cannot switch to WLAN because of pending actions. In those cases, WLAN access is delayed to the next WiMAX UL subframe. If a VoIP packet is awaiting a transmission in WLAN, an additional delay of another WiMAX frame may be imposed but this also ensures a timely network entry in WiMAX.

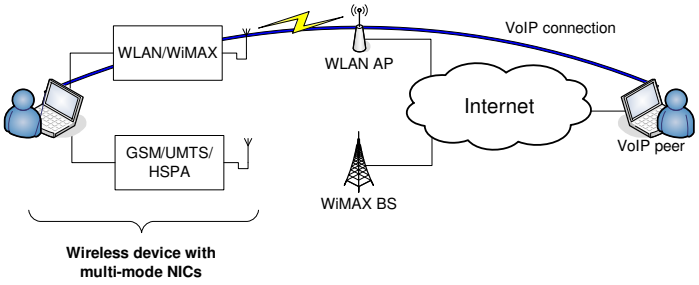


Figure 5.20: Network scenario

5.7.2 System and Problem Formulation

For our analysis, we refine our reference scenario from Sec. 4.4. As shown in Fig. 5.20, the device of the user is equipped with several multi-mode NICs, whereby WLAN and WiMAX share the same NIC with a dual-mode transceiver. We assume that our dual-mode receiver is able to change access technologies within insignificant time, thus following hardware designs being able to switch at "a single clock cycle" [159].

For WLAN, we focussed on 802.11g ERP OFDM with data rates ranging from 6 to 54 Mbps. For WiMAX, we relied on the *WirelessMAN-OFDMA PHY* [143], the TDD mode and the parameters of the *mobile profiles* specified by the WiMAX Forum [160]. As shown in Fig. 5.19, the WiMAX TDD mode introduces periodic frames, each consisting of a down- and an uplink subframe. Depending on the selected WiMAX PHY, the actual frame format changes. We shortly review the basic frame structure, highlighting the relevant parts for our analysis. For the WirelessMAN-OFDMA PHY, a *preamble* for synchronization starts the downlink subframe. Directly afterwards, the *frame control header (FCH)* and the *downlink map (DL-MAP)* message are transmitted. The FCH specifies the OFDMA sub channels used for the transmission of the DL-MAP message together with its length, while the DL-MAP defines the structure of the remaining downlink subframe consisting of so-called *bursts*. The DL-MAP message is followed by the first *DL burst* which includes *uplink map (UL-MAP)* and *DL/UL Channel Descriptor (DCD, UCD)* messages. Similar to DL-MAP for the downlink, UL-MAP gives the structure of the uplink subframe. DCD/UCD messages include information for receivers to decode a burst, such as the applied MCSs. FCH, MAPs, and DCD/UCD messages are encoded with the MCS QPSK-1/2.

Via the WLAN access cell, the user has an active VoIP communication session (G.711 voice codec with 20 ms packetization). Thus, the joint WLAN/WiMAX transceiver is blocked, such that no simultaneous access to WiMAX may be possible. The VoIP call has the usual strict QoS constraints (maximum packet loss of 5 percent and an end-to-end delay of 150 ms at most, cf. [161,162]). Several reasons exist for lost or delayed VoIP packets. They may stem from the 802.11 access cell or the wired part, e.g., a DSL provider, the Internet side or any combination of the involved entities. Even the wired part can significantly contribute to variations in the VoIP QoS. Measurements over backbone net-

Table 5.2: Parameters according to IEEE 802.16e and WiMAX Forum

frame duration [ms]	T_{frame}	5
bandwidth [MHz]	BW	3.5, 5, 7, 8.75, 10
cyclic prefix ratio	G	1/8
FFT size	N_{FFT}	512 (3.5, 5 MHz), 1024 else
sampling factor	n	28/25 (5, 10 MHz), 8/7 else
#PUSC subchannels	$N_{\text{DL-PUSC}}$	15 (512 FFT), 30 (1024 FFT)
#FUSC subchannels	$N_{\text{DL-FUSC}}$	8 (512 FFT), 16 (1024 FFT)
code rate	c	1/2, 2/3, 3/4
modulation level	m	2, 4, 6
modulation		QPSK, 16QAM & 64QAM
number of contiguous	S_{preamble}	1
symbols in time	$S_{\text{FCH, DL-MAP}}$	2
DL burst #1 [Byte]	L_{burst_1}	301
DL burst #2 [Byte]	L_{burst_2}	163

works have shown a temporal diverse behavior between different paths, whereby some of them even exhibit reoccurring patterns with respect to temporal higher delays [163]. As a basis of this work, we assume that the end user with the wireless device perceives some variations in the VoIP QoS due to jittering delay in the wired part. Although they do not bring the overall QoS below the acceptable level immediately, the user has perceived a slight degradation. To circumvent a potential stronger impairment, a handover is imminent, although the wireless device itself perceives a good WLAN channel.

For this scenario, we consider a solution that alternates the ongoing VoIP over WLAN communication with the network entry in WiMAX, which is the basis for the establishment of another path to the VoIP peer. In order to allow a fast setup of the alternative path, the WiMAX network entry has to be conducted as quickly as possible. Thereby, the fundamental question appears how to support additionally the VoIP communication over the WLAN path in a standard-compliant way.

5.7.3 Analysis of Timing for WLAN/WiMAX

The duration of the WiMAX DL subframe limits the available time for VoIP transmissions in WLAN and vice versa. The analysis takes into account the maximum duration of communication patterns in each technology, such that the wireless device can be still present for WiMAX DL subframes and can transmit VoIP without any quality distortions in WLAN. For this, we assume an idle channel in WLAN and as a starting point

Table 5.3: $t_{\mathrm{WLAN\,max}}$ (ms) for 802.11g ERP OFDM

(Mbit/s)	Null-data	1 VoIP packet in DL	1 VoIP packet in UL	1 VoIP packet in UL & DL each
54	1.23	1.39	1.26	1.41
6	1.35	1.81	1.62	2.10

no other active device in WiMAX. Later in Sec. 5.7.6 we show results for various DL loads in WiMAX.

IEEE 802.16e Timing Issues: The duration of the DL part takes its maximum for the network entry if DL-/UL-MAP, UCD, DCD (within DL-burst #1) and (the largest) RSP message (DL-burst #2) are transmitted together in one DL subframe. Eq. 5.5 states the duration of the DL subframe in this case:

$$t_{\text{WiMAX-DL}} = \Delta t_{\text{symbol}} \{S_{\text{preamble}} + S_{\text{FCH, DL-MAP}} + S_{\text{DL-PUSC}} + S_{\text{DL-FUSC}}\}. \tag{5.5}$$

There, the sum of S_{preamble}, $S_{\text{FCH, DL-MAP}}$, and $S_{\text{DL-PUSC/FUSC}}$ gives the number of contiguous symbols in the time domain, whereby Δt_{symbol} denotes the duration of one symbol. Due to OFDMA, symbols can be allocated in time as well as on different sub channels. In WiMAX notation, a *slot* either consists of a single symbol or of two symbols (subsequently in time) on a sub channel denoted as *fully or partially used sub-carriers (FUSC, PUSC)*. Again, we are interested in the overall duration of the messages inside a downlink subframe, thus we derive the resulting duration from the applied slot to symbol allocation in time. For this, we consider the number of occupied slots N_{slot}, the resulting number of contiguous symbols in time, and the symbol duration Δt_{symbol} as given below, whereby we apply FUSC for the transmission of the two DL-bursts:

$$S_{\text{DL-PUSC}} = 2 \left\lceil \frac{N_{\text{slot}}}{N_{\text{DL-PUSC}}} \right\rceil \text{ [symbols]},$$

$$S_{\text{DL-FUSC}} = \left\lceil \frac{N_{\text{slot}}}{N_{\text{DL-FUSC}}} \right\rceil \text{ [symbols]},$$

$$N_{\text{slot}} = \left\lceil \frac{L_{\text{burst}_X}\,[\text{Byte}] \cdot 8\,[\text{bit/Byte}]}{c \cdot m\,[\text{bit/data-sc}] \cdot 48\,[\text{data-sc/slot}]} \right\rceil,$$

$$\Delta t_{\text{symbol}} = (1+G)\frac{N_{\text{FFT}}}{n \cdot \text{BW}}.$$

Table 5.2 summarizes selected parameters and their values according to IEEE 802.16e OFDMA [143] and the mobile profiles from the specification of the WiMAX Forum [160].

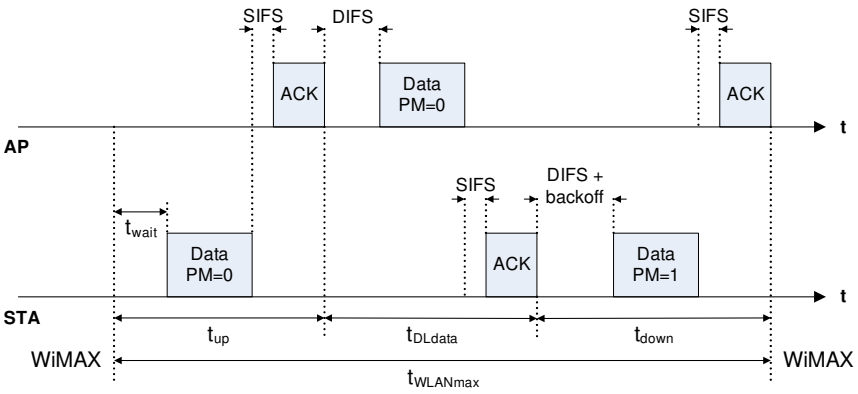


Figure 5.21: WLAN PS signaling with up- and downlink transmission

802.11 Timing Issues: The maximum duration of VoIP transmissions in WLAN occurs, when AP and STA apply the most robust data rate. Depending on the situation whether packets are awaiting their transmission in UL as well as DL or in one of the directions only, the PS signaling and its duration changes.

Fig. 5.21 shows the worst case, which consists of the wakeup, the exchange of one VoIP packet in UL and DL, and finally the sleep signaling. For this, Eq. 5.6 specifies the maximum active duration in WLAN:

$$t_{\mathrm{WLAN\,max}} = t_{\mathrm{up}} + t_{\mathrm{DLdata}} + t_{\mathrm{down}}, \tag{5.6}$$

where the individual components are determined as follows:

$$t_{\mathrm{up}} = t_{\mathrm{wait}} + t_{\mathrm{SIFS}} + t_{\mathrm{ACK}} + \begin{cases} t_{\mathrm{VoIP}} & \text{pkt in UL,} \\ t_{\mathrm{Null}} & \text{else,} \end{cases}$$

$$t_{\mathrm{DLdata}} = \begin{cases} 0 & \text{no pkt in DL,} \\ t_{\mathrm{DIFS}} + t_{\mathrm{VoIP}} + t_{\mathrm{SIFS}} + t_{\mathrm{ACK}} & \text{else,} \end{cases}$$

$$t_{\mathrm{down}} = t_{\mathrm{DIFS}} + \mathrm{rand}_{\mathrm{uniform}}(0, \mathrm{CW}_{\mathrm{min}}) \cdot t_{\mathrm{slot}} + t_{\mathrm{Null}} + t_{\mathrm{SIFS}} + t_{\mathrm{ACK}}.$$

Table 5.3 gives the maximum active communication duration for 802.11g ERP OFDM for the cases of no traffic, a packet in each direction only, and for both up- and downlink. For the cases with present VoIP traffic, the highest values for the most robust MCS with 6 Mbps have been selected as thresholds (highlighted in grey).

5.7.4 Duration of WiMAX Neighbor Discovery

The first step of the heterogeneous opportunistic approach tackles the discovery of a neighboring WiMAX network. Hereby, we shortly derive constraints for the selection of the scanning interval. As the 802.16e standard [164] in principle allows for WiMAX

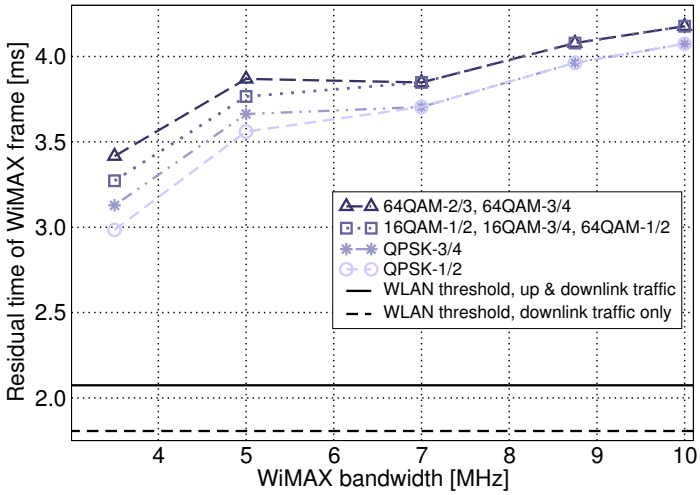


Figure 5.22: Available residual time of WiMAX frame

frames size up to 20 ms, we discuss the neighbor discovery aspect for the complete range of frame sizes. On the one hand, maximizing the scanning duration speeds up the WiMAX discovery process, i.e., to minimize the number of required scan attempts. On the other hand, we propose to stay below the packetization interval for the considered VoIP traffic, which in this case is 20 ms, in order to not induce large additional delays. Moreover, the opportunistic interval must not be equal to one or to multiples of the WiMAX frame sizes, since then the periodicity leads to problems in finding the other network. Back in Sec. 5.5.3, we already discussed that opportunistic intervals with prime numbers fulfill this last requirement. Overall, this leads to an optimal opportunistic interval size of 19 ms. With the choice of this interval value, a WiMAX network in a specific frequency band will be found in just one opportunistic interval if the WiMAX frame size (T_{WiMAX}) is equal or smaller than 12.5 ms, which is the case for mobile WiMAX with of 5 ms frames. With T_{WiMAX} of 20 ms, Eq. 5.4 holds and results in a maximum of 2, 3, and 4 scan attempts with 5 percent probability each, while just a single attempt is required in 85 percent. Thus, compared to the task of finding another WLAN AP (cf. Sec. 5.5.3) the number of required scan attempts for WiMAX is very low and can be seen as an uncritical part of the network entry.

5.7.5 Single Device: Feasible Parameter Space for Mobile WiMAX

Once the neighbor discovery has been completed, all subsequent steps of our approach require the device to stay for the DL subframes within the WiMAX cell (cf. Section 5.7.1).

When the device stays in WiMAX for the duration of the DL subframe, a residual duration occurs, which is analyzed in the following. Since we assume that this residual

Table 5.4: Maximum WiMAX downlink load (in percent of frame size)

$t_{\mathrm{WLAN_{max}}}$	**WiMAX frame duration [ms]**						
	2.5	4	5	8	10	12.5	20
1.81 ms	27.8	54.9	63.9	77.4	81.9	85.6	91.0
2.10 ms	17.0	48.2	58.5	74.1	79.3	83.4	89.6

time span is used for WLAN communication, t_{residual} has to be greater than the WLAN thresholds defined in Section 5.7.3:

$$t_{\mathrm{residual}} = T_{\mathrm{frame}} - t_{\mathrm{WiMAX\text{-}DL}} > t_{\mathrm{WLAN_{max}}}. \tag{5.7}$$

The residual time values were calculated for all combinations of channel bandwidths and MCSs (for DL-burst #2) listed in Table 5.2. Fig. 5.22 shows the results for the case of no other DL-load in WiMAX: the residual time of the WiMAX frame stays far above the WLAN thresholds for all parameter combinations. Thus, our approach is feasible if no other traffic is present.

5.7.6 Multi-Device Case: Load Dependency

The last part deals with the influence of background traffic in the WiMAX DL subframe, i.e., when the BS serves also other devices. This further reduces the time span that is utilized to switch to WLAN. We identify the performance limits for this duration as a function of present traffic load in the WiMAX DL subframe.

We define the maximum WiMAX DL load L_{max} as fraction of T_{frame}, under which the timing constraints of our solution still work:

$$\mathrm{L_{max}} = 1 - \frac{t_{\mathrm{WLAN_{max}}}}{T_{\mathrm{frame}}}. \tag{5.8}$$

Table 5.4 gives again results for various WiMAX frame sizes and both WLAN thresholds. If T_{frame} is far below the VoIP packetization interval, it is pretty likely that there is only one packet waiting in UL or DL. In this case, the smaller WLAN-threshold applies. For larger T_{frame} above 10 ms, the second WLAN threshold is likely.

Overall, for 802.16e OFDMA with 5 ms frames, our solution is applicable if the DL part consumes not more than 63.9 percent (or 3.2 ms) of the frame duration. Fig. 5.23 finally connects the results with and without other DL traffic graphically: the upper edge of the inclined plane represents the case with no other background traffic (and QPSK-1/2 MCS for all messages). Dependent on the WiMAX parameter combination, there are 18.1 percent (3.5 MHz bandwidth) and 39.9 percent (10 MHz) of the WiMAX frame remaining for the DL load.

5.7.7 Summary of Results

This section presented a sample application of our opportunistic approach incorporating the first design flavor. The evaluation of the scheme highlighted the timing constraints

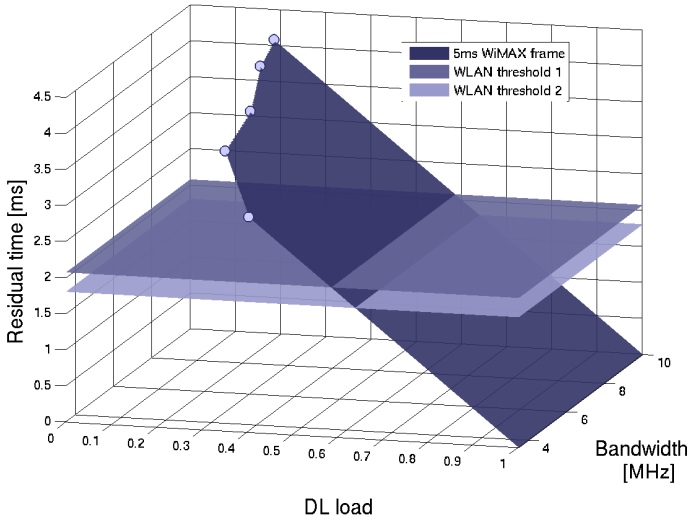


Figure 5.23: Residual time with different levels of mobile WiMAX DL load

for WLAN and WiMAX under the assumption that the WLAN channel only transports VoIP and power-save traffic of the opportunistic device. Our results identify the limits for different mobile profiles of the WiMAX forum such that QoS constraints of real-time VoIP traffic in WLAN are not violated. These limits are intended to support simple decisions on device side when our scheme may be applied in principle. With this, we effectively avoid that end devices try to rely on our approach for WiMAX parameter combinations in which QoS constraints may be violated. Finally, we shall emphasize that this analysis has been conducted for a tight timing approach, where we give strict priority to the WiMAX network entry. As such, the obtained limits represent an upper bound and may be relaxed for other priority schemes. Nevertheless, for other applications, one may further tradeoff the priority of WiMAX timing and WLAN channel access as detailed in Sec. 5.2.1. In case of strongly fluctuating WLAN channels, for example, it may be important to give more channel access time to WLAN (i.e., prioritize the transport of VoIP data) and postpone subsequent steps of the WiMAX network entry process (in the valid range of the IEEE 802.16 standard) instead.

5.8 Final Remarks: The Impact of the Queuing Policy on AP Side

Broadly speaking, 802.11 power save is known to work best together with idle up to modest load conditions. A high network load together with PS STAs may play a critical role in terms of additional packet delay, wasted medium resources, and high energy expenditures. More specifically, the dominant issue lies in the way how the 802.11 AP

handles temporarily buffered data packets for a STA being in PS mode. As such strategies are beyond the scope of the 802.11 standard, different vendor-dependent solutions have evolved over the last decade. Rozner et al. [152] showed that WLAN APs fall in two categories regarding their strategies handling buffered frames for PS STAs. The authors denote these categories as "normal" and "high priority scheduling". In both, the AP usually has separate queues for PS STAs, in which arriving downlink traffic is buffered while STAs are in doze state. However, once the AP is triggered to conduct downlink transmissions to the STA, the behavior varies between the strategies. APs following the "normal" scheme dequeue the data packets destined to this STA from the corresponding PS queue and insert them *at the end* of the FIFO queue which is used for all outgoing downlink traffic. Thus, the downlink packets which were buffered already for power-save reasons may perceive again an additional delay, while the STA has to stay in the awake state for this time span. Obviously, the higher the downlink network load is, the more packets wait in the FIFO queue of the AP and the stronger this tendency amplifies. In contrast, for the "high priority" approach, 802.11 APs insert all packets to our STA in a separate queue being flushed prior the FIFO queue for traffic of non-PS STAs. While with this approach, downlink packets to PS STAs are sent out timely, it may lead to an unbalanced treatment of the other (non-PS) STAs in the WLAN cell. Note that some (more recent) APs being 802.11n-capable [6] have been shown to apply the "high priority" scheme [152]. Also, in the same paper, a sophisticated approach denoted as "fair scheduling" was presented aiming at a balanced handling of power save as well as active mode STAs, whereby fairness is considered regarding the delivery of downlink packets based on their previous, sequential arrival at AP.

Generally, we suggest to use the opportunistic approach with the 802.11e enhancements for prioritized medium access. These capabilities are announced by an AP within its beacon frames. With an AP supporting these features, the scheduling issue diminishes as 802.11e introduces four priority queues allowing a separated handling of different traffic types. This avoids a single queue at AP, which buffers all downlink traffic in a FIFO manner. Accordingly, we designed the opportunistic approach such that it works well with APs employing "fair" or "high priority scheduling" strategies according to the definition of Rozner et al. If one may want to apply the opportunistic approach nevertheless with "normal" scheduling strategies at AP, we suggest to initially 'test' the network regarding an applicability of our opportunistic approach, especially in loaded scenarios where a significant queuing delay might appear. For such an initial test, we propose to analyze how long it takes the AP to deliver a data frame to the STA. In order to let the opportunistic approach work smoothly, we can derive a time interval $\Delta t_{\text{timeout}}$, in which the data frame must be received. Thereby $\Delta t_{\text{timeout}}$ is just a certain fraction of the periodic opportunistic interval with which we alternate doze and awake states of a STA. If within $\Delta t_{\text{timeout}}$ a downlink frame is not received by the STA, we conclude that the given AP under the current load condition is not capable to support the opportunistic approach.

Chapter 6

Handover Candidate Selection

In Chapters 3 and 4, we have discussed that offloading data traffic to 802.11 hotspots is one of the key techniques with which cellular operators try to manage the increasing traffic demand of their users in WWANs. This offloading is simply conducted today once an end-device has WLAN connectivity. By this, offloading neglects the viewpoint of the WLAN hotspots, whose overall performance in terms of the aggregated throughput highly depends on a 'proper' mix of devices. Under such a 'proper mix', we understand that a WLAN hotspot should only serve devices which do not impact negatively on the performance of the whole cell. The option for such 'bad' devices already being served by a hotspot actually is the 'onloading' back to WWANs, leading to free resources in the WLAN that may be better utilized by other 'offloading' candidates instead.

We present a performance metric for a selection of devices in a WLAN hotspot subject to WWAN onloading, thereby considering the peculiarities of 802.11 networks. More specifically, we aim to maximize the efficiency regarding WLAN resources occupied by data transmissions, thus allowing to serve either more devices or to enable an increase in layer 2 throughput. We base our definition of efficiency on how each device exploits its share of WLAN resources in terms of the channel occupation time together with all the MAC overhead as result of contention, interference, and fluctuating channels.

The chapter is structured as follows. First, we refine our underlying scenario and discuss related assumptions (Sec. 6.1). Second, we define the scope of our onloading decisions (Sec. 6.2). Afterwards, we present the design of our selection metric (Sec. 6.3) and analyze its effects by an intuitive application example (Sec. 6.4). Further, we discuss our selection metric in the context of an onloading decision scheme (Sec. 6.5), and present a comprehensive performance evaluation by comparing our solution with approaches from the literature (Sec 6.6). Finally, our remarks at the end of this chapter (Sec. 6.7) discuss how the different measures for our selection scheme are obtainable from off-the-shelf WLAN devices.

The work of this chapter was published before; the initial design of our onloading metric together with a selected test case was included in [165, 166], the extensive performance evaluation of our scheme was published in [167].

6.1 Refined Scenario and Assumptions

Considering our reference scenario from Sec. 4.4, we focus in this chapter on a single WLAN hotspot, which applies the architectural framework given in Sec. 4.5. Thereby, the area of the WLAN hotspot is covered by a WWAN with 3GPP 3G/4G technology. Again, all end users are static and have multi-standard devices supporting both wireless accesses, WLAN as well as WWAN. Further, each active end device has an on-going traffic stream in up- and downlink direction, both either via WLAN or WWAN. In addition, we assume that the number of active end devices within the coverage area of the 802.11 hotspot exceeds the WLAN capacity in terms of accomplishable devices by far. To handle high-load situations in WLAN, we presume that an admission control scheme on WLAN side regulates access for devices that are going to be offloaded from WWAN to WLAN (compare Sec. 4.3.4). Without focussing on a specific admission control algorithm, we just assume that it fulfills the following condition: It grants only access to associating devices, if the hotspot still has free resources in terms of channel occupation time, such that a target minimum QoS level can still be supported for the vast group of on-going traffic streams in WLAN. Nevertheless, we note that a minimum QoS level for individual traffic streams may still be violated, if the corresponding devices suffer from frequent transmission errors due to bad channel conditions. Regarding end users currently not being served by WLAN, we assume that they are nevertheless accommodated by the WWAN. The precise number of end users with their devices and the type of traffic patterns varies between our simulation studies and is detailed below in each corresponding section.

6.2 Scope of Decisions for Handovers from WLAN to WWAN

A handover from WLAN to WWAN in the above scenario emerges in two different contexts, regarding QoS as well as resource management aspects. First, the QoS of a traffic stream in terms of throughput, packet delay or packet losses may degrade for a WLAN device even below a targeted minimum level, e.g., due to rapid changes in the radio link quality. In addition, even the minimum QoS level in a WLAN hotspot may not suffice the requirements of an end-user. In the literature, various layer 2 indicators for such a *QoS-centric handover* were proposed ranging from RSSI, to packet losses, retransmission and data rate measures. We assume throughout this chapter that the device itself recognizes the need for such a handover and conducts required actions accordingly.

In contrast, the second type of handovers aim for a resource management in the WLAN part of our heterogeneous network. We refer to the second type as *forced handovers* to WWAN. On the WLAN network side, in our selected architecture the RMC (compare Sec. 4.5), decides and triggers such a handover for a selected device. Exactly for the type of forced handovers, this chapter discusses a novel criterion analyzing how we get most out of the WLAN hotspot in the context of WWAN onloading strategies considering the peculiarities of 802.11 networks. Once an end device is associated in WLAN[1] and runs an ongoing traffic stream, it usually sticks to its AP as far as no

[1] we refer to this as a WLAN STA in the following

need for a QoS-centric handover emerges. This may become more and more problematic, when the WLAN hotspot is loaded up by an admission control scheme (compare Sec. 4.3.4) close to its maximum. At this operational point of the WLAN cell, new arriving STAs may be more efficient in terms of occupied resources than already present STAs. In other words, it may be advantageous to have a means that selects 'bad' STAs being candidates for an onloading back to WWANs, thus allowing larger amounts of data to be accommodated in the WLAN hotspot.

6.2.1 Assessing Occupied WLAN Resources

Previous work defined occupied resources as the amount of time that the WLAN channel was busy in relation to the measurement duration, being denoted as "channel utilization" [92]. More precisely, the latter is defined "as the percentage of time that the AP sensed the medium was busy, as indicated by either the physical or virtual carrier sense (CS) mechanism" [92]. For a load-balancing of a WLAN/cellular network, Song et al. [168] already considered the channel utilization of 802.11 as a load measure. Further, in different areas of WLAN research, e.g., regarding admission control, load balancing, and radio resource management, such a load measure based on the channel occupation time has been commonly applied to define the amount of resources required to serve an associated STA, e.g., in references [8,169–171]. Following previous works, we apply such a measure regarding the occupied channel time per WLAN device that is commonly also denoted as *airtime*. Throughout this thesis, we define airtime as the sum of time spans spent for wireless transmissions of a given device. We give its exact computation further below in Sec. 6.3.1.

In the following we shortly discuss why occupied WLAN resources are commonly considered on the basis of airtime measures. In a WLAN cell, occupied resources cannot simply be determined just by accounting transmitted bits per second. In particular, the achievable aggregated throughput of a hotspot highly depends on the number of contending STAs, STAs' positions, the channel quality, the transported traffic patterns, and on the way how WLAN STAs adapt their data rates. First, in WLANs, all senders contend for the medium access per data frame thus being independent of the actual frame size. 802.11, besides the duration for the channel contention, introduces an overhead per data frame in terms of framing, inter-frame spaces, and immediate acknowledgment (cf. Chapter 2). As a result, the smaller the payload of transmitted data frames, the smaller becomes the useful fraction of time being applied for the actual payload, while the duration of the overhead remains constant in contrast. It is well-known in the area of WLANs that this relation has a strong impact on the network capacity, e.g., in terms of throughput, which lowers with decreasing payload sizes. Second, the traffic patterns of STAs by their nature diverge regarding the frequency as well as in the size of transmitted data frames. As a result, the actual mix of traffic streams in a WLAN cell determines the *load level* as well as the *collision level*. The latter depends on the number of WLAN STAs contending for medium access at a given instance of time. Third, the different positions of WLAN STAs also have an impact on the *received signal strength* with which frames arrive at the receiver as a result of path loss. Due to time varying channels as

well as interference, the signal strength further obeys a stochastic process leading to retransmissions of corrupted data frames additionally increasing the load as well as the collision level in the cell. Last but not least, each STA selects the PHY data rate for its transmissions on some proprietary heuristics. In a nutshell, all these aspects impact the *duration* for which a WLAN STA occupies the channel. Thus, approaches with throughput-based metrics, e.g., estimates regarding the remaining capacity as frequently considered in the context of heterogeneous handover decisions (compare Sec. 3.4.3), are not well suited to assess the amount of occupied resources in a WLAN hotspot serving a mix of STAs with different traffic patterns.

6.2.2 Objectives for Onloading Handover Decisions

Considering the airtime for each device, we aim to select devices for an onloading handover from WLAN back to WWAN. Thereby, we target to minimize occupied resources of associated STAs, allowing us either to serve additional end users or to enable a higher throughput for associated devices in the WLAN hotspot. We aim to minimize occupied resources by keeping STAs inside the hotspot operating close to a minimum amount of airtime for given type of traffic. For this, we tackle to identify devices for an onloading that rapidly contribute to the present load level in the WLAN just because of strong expenses in terms of framing, frequent retransmissions and a usage of low-rate MCSs. We denote these STAs as *bad* devices throughout this chapter.

For a given device, the airtime identifies only its contribution to the load of a WLAN hotspot. Still, by comparing airtime measures for WLAN STAs with *different* types of traffic streams, one can a priori not identify a 'bad' device. Let us consider the basic reason for this in more detail by a simple example consisting of a WLAN hotspot with two end-user devices, one running a high-bit rate video traffic stream, the other one conducting a VoIP call resulting in low traffic rates. As the video devices has more data to transmit than the VoIP device, also the amount of occupied WLAN resources (the airtime) is higher. Nevertheless, the airtime does not give us an indication in this case, how each device has been exploiting the occupied resources. However, exactly this is the crucial point for the 802.11 technology in which devices contend for the medium access and, in addition, link rate adaptation and error recovery mechanisms such as retransmissions further increase the airtime for a given STA. In summary, from the above discussion, we identify two issues for the handover candidate selection:

A. What minimum amount of airtime is obligatory for the transport of a specific traffic stream, and what part, which we denote as *surcharge*, is essential because of current conditions such as load, collision and interference level, as well as the position of the device?

B. For a given traffic stream with a certain data packet size, how much *overhead* in terms of framing, inter-frame spaces, and immediate ACKs is introduced by 802.11?

For the sake of completeness, we point out that B) reflects the relation of the overhead to the total duration of a transmission as discussed in Sec. 6.2.1. Considering issues A) and B), we aim for a bare comparison of WLAN STAs regarding how efficient each is using its amount of airtime. Again, we tackle to identify devices with their traffic that contribute to the present load level the WLAN access network but benefit only marginally from these expenditures. We denote such behavior as *inefficiency* in the following.

6.2.3 Challenges for a Practical Design

In the context of the above objectives, we further identify four fundamental challenges from the layer 2 perspective to allow the selection of devices being candidates for an onloading with off-the-shelf WLAN equipment:

1. A clear and easy comparison of WLAN STAs shall support decisions about an identification of 'bad' devices which may be subject for a handover back to WWAN technology.

2. The layer 2 WLAN parameters considered for such a decision should be easily measurable on STA and AP side.

3. 802.11 networks offer a zoo of technology-specific parameters each reflecting a certain aspect of the hotspot. As a result, selected parameters shall be combined to a unified, easily computable performance metric.

4. Certain devices may not cooperate with a hotspot in the sense that they do not want or are not able to support the participation in layer 2 specific measurements. Thus, our approach still needs to be operable for non-cooperative devices.

6.3 Design of the Selection Metric

Following the objectives in Sec. 6.2.2 and the challenges in Sec. 6.2.3, we designed a unified performance metric on the basis of the airtime of each STA. We denote it as the *inefficiency metric*, as we aim to select STAs for an onloading to WWAN, which are not efficient in terms of using their resources. The inefficiency metric consists of two parts, namely the *surcharge* and the *overhead factor*. Basically, the surcharge tackles objective A), while the overhead factor handles our goal B) from Sec. 6.2.2.

We constrain our basic definition of the inefficiency metric to an evaluation of the airtime spent for a given STA. Note that by this we do not rate the actual load level that the STA generated. In other words, we present a metric reflecting the efficiency regarding the usage of airtime, independent of whether the actual airtime value is high or low. Later, in Sec. 6.5, we relax this constraint by considering handover decisions that base both, on our metric and the absolute airtime (as a measure for the load contribution) of a device.

In the following, we first define both parts of the inefficiency metric, surcharge and overhead factor, and afterwards describe, how we combine these measures to our unique performance measure.

6.3.1 Surcharge: Reflecting the Efficiency of Wireless Transmissions

Let us start the discussion for our surcharge part from a simple point of view: If we know the lowest costs for transmissions on MAC level, we can further compare it with the actual appearing costs. For this, we first need to derive the minimum costs which we define as the minimum airtime for a given traffic stream to be transmitted according to the regulations of the 802.11 standard. Note that the absolute smallest possible airtime for the transport of a traffic stream appears when one assumes ideal conditions (i.e., no path loss, fading, interference, packet collisions, packet errors, etc.) allowing transmissions to occur at the highest link data rate and without any retransmissions. Obviously, this evokes the lowest possible load on the channel for a given set of transmissions. In contrast, all means for error control and adaptation to channel conditions (e.g., rate adaptation, retransmissions) lead to an increase of the load on the wireless channel in real systems. This increase of channel load in relation to the lowest possible load is the first measure which reflects the additional expenditures required to deal with the real conditions inside a hotspot. We derive this part from the basic definition of efficiency [172]: In engineering, efficiency is usually defined as the relation of system's output ϑ to the overall effort ψ:

$$\eta = \frac{output}{effort} = \frac{\vartheta}{\psi} \tag{6.1}$$

Efficiency η can range in the interval $[0, 1]$, whereby effort values much larger than output values ($\psi >> \vartheta$) lead asymptotically towards efficiency values of zero. The design rationale behind this part is to identify STAs with smallest efficiency values as handover candidates. However, it may be difficult to distinguish between two small efficiency values close to zero although the corresponding difference of effort values may be significant. Hence, the reciprocal is applied to enable comparability.

$$surcharge\, \zeta = \eta^{-1} = \frac{\psi}{\vartheta} \tag{6.2}$$

In summary, the surcharge increases the more airtime a WLAN device consumes for its type of data transmissions. In other words, the higher the surcharge is, the more channel time a STA occupies as a result of transmissions at low link data rates or retransmissions.

Without loss of generality, in the following we discuss the parts of this measure for the transmission of a single data frame. The computation for a whole unidirectional stream of packets is straightforward as it requires a summation over the transmitted frames. For a single transmission of a MPDU in WLANs, the effort ψ depends on the state of the wireless channel, the choice of a modulation scheme, the collision level as well as the number of retransmissions. All these parts have an impact on the effort for a transmission in a way that they affect its duration. Thus, we use the duration for a complete transmission sequence in order to determine the effort required for the

transport of the MAC service data unit (MSDU) (Eq. 6.3). There, the number of trials represents the (re)transmissions that were required to ensure the delivery of the MSDU.

$$\psi = t_a = \sum_{i=0}^{\#\text{trials}} \Delta t_i \tag{6.3}$$

The airtime t_a, being equivalent to ψ, represents the amount of time that the wireless medium is occupied (or reserved, in case of inter-frame spaces and NAV settings)[2]. Thereby, the time Δt_i for each trial is computed as follows:

$$\Delta t_i = t_{\text{IFS}} + t_{\text{d}}(\text{Rate}_i) + t_{\text{ack}} \tag{6.4}$$

This includes the whole transmission sequence consisting of the inter-frame spaces DIFS or AIFS and SIFS (t_{IFS}), the duration t_{d} of the complete data frame 'on air', where the data part is encoded with a certain modulation scheme Rate_i, and the time span for the immediate acknowledgment t_{ack}.

In contrast, we define the system's output at MAC level ϑ as the absolute smallest possible duration for the whole transmission that would be required in case of an ideal error free channel (Eq. 6.5).

$$\vartheta = \Delta t_{\text{opt}} = t_{\text{IFS}} + t_{\text{d}}(\max \text{Rate}) + t_{\text{ack}} \tag{6.5}$$

This output definition includes the duration of the whole data frame when the data part is encoded with the highest MCS *max Rate* and a single transmission attempt is conducted. Thus it serves as a reference case that implies the smallest possible effort.

6.3.2 Overhead Factor: Penalizing Short Frames

While the surcharge is a measure for the inefficiency regarding the transmission of MPDUs, it does not tell anything about the suitability of WLANs to transport these frames with their specific size. To cover our objective B) from Sec. 6.2.2, we introduce the overhead factor in the following.

As discussed in Chapter 2, 802.11 introduces a certain amount of overhead (PHY framing, inter-frame spaces and immediate ACK) for one transmission regardless of the MSDU size. Thus, the smaller the MSDU size, the more contributes this overhead to the channel load. In other words, 802.11 becomes less optimally utilized (cf. Sec. 6.2.1).

To accommodate for this behavior, we compare the duration of the overhead with the duration of the (optimal) time for a complete transmission. In other words, we identify the fraction of overhead for the corresponding transmission as follows

$$\alpha = \frac{\Delta t_{\text{oh}}}{\Delta t_{\text{opt}}} = \frac{\Delta t_{\text{oh}}}{\Delta t_{\text{MSDU}_{\text{opt}}} + \Delta t_{\text{oh}}}. \tag{6.6}$$

[2]We do not include backoffs in the calculation of the airtime, as we aim to consider the duration for which the WLAN channel is occupied. In other words, we are not interested in the time span for which a STA contends for channel access.

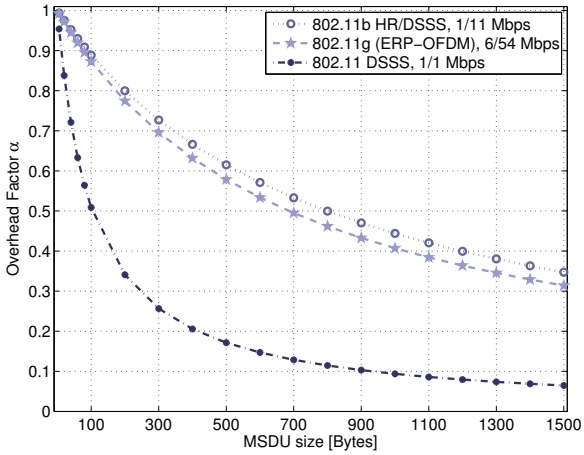


Figure 6.1: Overhead factors of 802.11/b/g PHYs

Here Δt_{opt} is again the smallest possible duration for a frame exchange from Eq. 6.5, $\Delta t_{\text{MSDU}_{\text{opt}}}$ represents the duration of the bare MSDU assuming the highest modulation scheme. Δt_{oh} includes all necessary overheads due to framing, inter-frame spaces, and immediate ACK. Further substituting $\Delta t_{\text{oh}} = \Delta t_{\text{opt}} - \Delta t_{\text{MSDU}_{\text{opt}}}$ in Eq. 6.6 leads to a more intuitive expression

$$\alpha = 1 - \frac{\Delta t_{\text{MSDU}_{\text{opt}}}}{\Delta t_{\text{MSDU}_{\text{opt}}} + \Delta t_{\text{oh}}} = 1 - \frac{\Delta t_{\text{MSDU}_{\text{opt}}}}{\Delta t_{\text{opt}}}. \tag{6.7}$$

The overhead factor starts for small MSDU sizes shortly beneath one and continuously decreases with larger MSDU sizes. Figure 6.1 displays the overhead factors for different MSDU sizes and three different 802.11 PHYs. The overhead curves for all three PHYs are monotonically decreasing with the size of the data part of the MAC frames. Thus, the higher the overhead factor is, the stronger is the penalty. Let us compare the curves for the 802.11 DSSS and the 802.11b HR/DSSS PHYs. Both have the same basic rate of 1 Mbps. However, 802.11b can transmit the data part of frames with the highest rate of 11 Mbps. Especially for small MSDU sizes being typical for VoIP traffic (e.g., 200 Bytes in case of G.711-coded speech and a packetization of 20 ms), we can see that the highest link data rate has a strong impact on the overhead factor. Generally, the higher the maximum link data rate, the smaller becomes the data part in time and as a result, the stronger is the impact on the overhead. Further taking 802.11g into consideration, one can see that the time duration of the total overhead values of 802.11g ERP OFDM PHYs are slightly lower than the 802.11b curve. This results from the fact that 802.11g ERP OFDM comes up with smaller slot times, shorter PLCP preamble and header as well as a higher basic rate of 6 Mbps—as compared to the basic rate of 1 Mbps for 802.11b HR/DSSS.

6.3.3 Composition to the Inefficiency Metric

In order to allow a handover candidate selection among users with heterogeneous traffic patterns, we combine the overhead factor α and surcharge ζ to a single inefficiency metric. We use the overhead factor thereby as a penalty for small transmitted data frames thus applying the product between both measures which we denote as *inefficiency metric*:

$$D = \alpha\,\zeta \tag{6.8}$$

So far, we have discussed the computation of the inefficiency metric and its parts in a simplified way by considering just a single data frame. Now we generalize it by summing up output ϑ and effort ψ over fixed-size intervals. With this summation, we enable to compute the metric value over multiple transmissions and larger time scales.

6.4 Demonstrating the Usability of the Inefficiency Metric

In the following, we compare the inefficiency metric with SNR-based decisions in scenarios, where the latter leads to optimal results regarding a maximum number of STAs in a WLAN hotspot. Under such conditions, both metrics—SNR as well as inefficiency metric—are expected to decide for the same STAs as handover candidates. By this, we show that we do not underperform compared to SNR decisions, even in scenarios where SNR-based criteria lead to optimal solutions.

As a show case, we select a simple but illustrative example in which it was intuitively clear which STAs are the 'bad' ones. For this testing, we concentrate on a scenario where all WLAN STAs apply the same traffic pattern and are distributed over the hotspot coverage area such that no hidden nodes appear. In such a setting, a selection of 'bad' STAs is very easy as the nodes with the highest distance to the AP have the worst channels, i.e., lowest SNR values. These STAs are the ones evoking the highest load level on the channel. Note that this is in line with the most common handover or access selection decisions in heterogeneous networks that are performed for the WLAN part on the basis of SNR values [59]. In such approaches, STAs with lowest SNR values are candidates for a handover. We will see that for *identical* traffic patterns on present WLAN STAs, SNR may be indeed an appropriate measure to judge the efficiency of transmissions.

This section shows that, under the above assumptions, decisions based on the inefficiency metric are identical to SNR-based handover triggers. Further, we give insights into the behavior of our metric for up- and downlink transmissions, reflecting the typical traffic asymmetry inside WLAN hotspots. Lastly, we identify the impact of handovers, both on the inefficiency and the QoS of on-going traffic streams.

6.4.1 Goals of Investigation

By means of simulations, we studied a WLAN hotspot scenario in which we identified 'bad' STAs being onloading candidates. Further, we analyzed the impact of our selection on users remaining in the WLAN cell, if a single candidate performed a handover from

WLAN to WWAN. Thirdly, we were interested in the impact of multiple handovers according to our approach: There, we handed over several candidates from WLAN to WWAN, while the same number of users (with the same service type) were put from WWAN to WLAN.

6.4.2 Set of Experiments

For the above goals, a set of three experiments was performed, which we denote as the *max. #nodes*, the *reduced*, and the *replaced set(s)*. The first experiment determines the maximum number of nodes such that the WLAN network is loaded (but not saturated) in a way that the QoS constraints of at least one node are violated. Second, we show the impact of a single handover from WLAN to WWAN when choosing the most 'inefficient' WLAN user. This experiment is called 'reduced set' since the total number of WLAN users decreases by one. In comparison to the maximum number of nodes, this experiment gives an idea about the approximate range of improvements due to the single handover of the most 'inefficient' user. Third, we studied the impact of multiple handovers according to our strategy. There, we conducted a replacement of nodes based on the results of the 'max. #node' experiment. Under replacement, we understand here that a node with a high metric value is triggered to perform a handover from WLAN to WWAN, while the WLAN network accommodates another node (either due to a handover from WWAN or a new, arriving user). Here, it is assumed that this new node is present near the AP with a distance of 10 m and represents the same user type as the one put from WLAN to WWAN. This third experiment was conducted with one to three replacements in total.

6.4.3 Simulation Scenario

In our simulations, we modeled the WLAN part of our scenario given in Sec. 6.1. The WLAN hotspot consists of an 802.11g AP that is 11e-capable by providing EDCA functionality. The EDCA model is described in detail in Appendix B.1. We assume to have VoIP users, which are equally distributed over the area of interest. We note again, that the exact number of VoIP users being active in parallel was determined by the max. #node experiment. The 802.11e/g parameters were chosen according to [6], leading to CW_{min} and CW_{max} of 3 and 7 for VoIP traffic.[3] To take into account that radio signals are not only affected by path loss but also by multi-path propagation, we applied our ns-2 simulation suite including a log-distance path loss, a Ricean fading, as well as an signal-to-interference and noise ratio (SINR) model as detailed in Appendices B.3 and B.4. Further, for the link data rate adaptation on AP and STAs, we use *adaptive auto rate fallback (AARF)* [173]. The AARF scheme is an extension of the *auto rate fallback (ARF)* algorithm, which increases the link data rate after ten acknowledged data transmissions and reduces its rate if two contiguous transmission attempts remain unsuccessful. As this static adaptation of ARF has been shown to be susceptible to fluctuating wireless channels, AARF dynamically tunes the number of contiguous suc-

[3] TXOPLimits were set to zero so that a single transmission per medium access attempt is performed.

cessful and non-successful transmissions for up- and downgrading the link data rates in addition to the ARF operation.

6.4.4 Node Placement and Traffic Model

In the simulation scenario, WLAN VoIP nodes are distributed equally over the area of interest, which has a shape of a quarter circle. There, the AP is located at the corner of the considered environment, such that no hidden nodes appear. All nodes have a VoIP call with a wired node outside the WLAN. The delay between the AP of the WLAN access cell and the wired nodes was set to 100 ms. All stations use an exponential ON/OFF model with mean ON and OFF durations of 1.004 s and 1.587 s [174]. During ON periods, voice packets are generated according to the G.711 codec [120] with a packetization of 20 ms, i.e., each voice flow has a 64 kbps peak rate with 160 Byte audio packets.

6.4.5 Metrics and QoS Constraints

For our simulation studies, we describe the selected metrics and the QoS constraints for the considered VoIP traffic in the following.

Surcharge We evaluated the inefficiency measure in order to visualize handover decisions. Since this study considers the same VoIP traffic pattern for all nodes, the overhead factor is identical for all. To keep it simple, we focussed on the surcharge part of the inefficiency metric in our simulation studies. For each transmitting STA, we determined its surcharge value over 100 ms during its active periods. All surcharge results were evaluated by batch means analysis [153] and mean values were plotted with their 95 percent confidence interval.

Application-Level Losses In order to assess the quality of the VoIP calls, we measured the loss of audio packets on application level over certain intervals. A loss can either occur due to lost or late packets. A packet is considered to be late if it arrives after a maximum network delay of 150 ms (similar to [161]) at the VoIP receiver so that it cannot be timely played out anymore.

QoS Constraints For each VoIP call, the quality should stay on an acceptable level. 'Acceptable' thereby means that a certain boundary for application level losses—consisting of packet losses and late packets—is not violated. With *packet loss concealment (PLC)* schemes and one-way delays up to 200 ms, random losses of up to 5 percent for G.711 are acceptable [175, p. 38, Fig. 29]. If five or more percent of the VoIP packets are lost, i.e., they have been dropped or they arrive with a network delay larger than 150 ms, the perceived quality is assumed to be temporarily low. We evaluated this criterion over intervals of four seconds such that we were able to analyze the incidence also of small periods with frequent (non-random) losses. The QoS boundary is defined as follows: If the perceived quality is temporarily low in 10 or more percent of the overall number of intervals, the quality degradation of the complete call is defined to be unacceptable.

Table 6.1: Uplink: quantiles at 5-percent packet loss

	Distance to AP [m]						
	14	49	83	116	134	144	149
max. #nodes	1.0	1.0	0.98	0.91	0.89	0.88	0.88
reduced set	1.0	1.0	0.99	0.95	0.93	0.93	—
1st replacement	1.0	1.0	0.98	0.93	0.90	0.90	—
2nd replacement	1.0	1.0	0.98	0.94	0.93	0.92	—
3rd replacement	1.0	1.0	0.99	0.95	0.94	0.94	—

Table 6.2: Downlink: quantiles at 5-percent packet loss

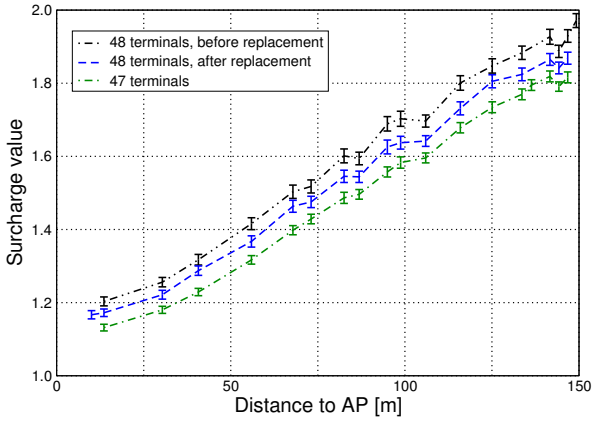
	Distance to AP [m]						
	14	49	83	116	134	144	149
max. #nodes	0.89	0.89	0.89	0.88	0.88	0.88	0.88
reduced set	0.94	0.94	0.94	0.93	0.93	0.93	—
1st replacement	0.91	0.91	0.91	0.90	0.90	0.90	—
2nd replacement	0.93	0.93	0.93	0.93	0.93	0.93	—
3rd replacement	0.95	0.94	0.94	0.94	0.95	0.94	—

6.4.6 Results

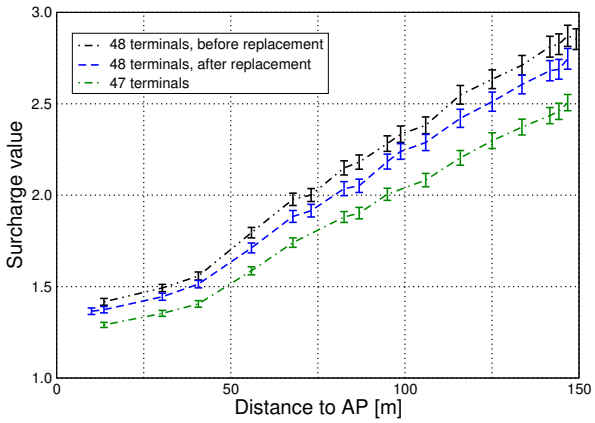
This section highlights that applying the surcharge metric in highly-loaded scenarios selects the 'bad' WLAN nodes, i.e., the devices which transmit at lower link data rates and which have to conduct a higher number of retransmissions than other STAs. First, we determine the maximum number of VoIP nodes and then show how handovers of 'bad' STAs improve the overall capacity of the cell in our selected scenario.

In the first experiment, we determined the maximum number of VoIP nodes in the WLAN cell. For this, we applied separate simulations, for which we increased the number of active VoIPs by one in each run. We continued with this until the QoS constraints of at least a single node were violated, thus obtaining the setting with a maximum number of VoIP calls. This is achieved with 48 VoIP nodes in total. Tables 6.1 and 6.2 show the cumulative probability of having five or less percent of application losses for all experiments in up- and downlink. While the QoS boundary is violated for all nodes in the downlink, the losses depend greatly on the distance between AP and STAs for the uplink, where boundaries are crossed for far nodes, only. This effect results from the asymmetric traffic distribution between AP and STAs and is discussed further below.

After identifying the operational point of the network where QoS constraints of several clients are violated, the second set of experiments shows the impact of a single handover. There, the handover candidate was selected according to our strategy of se-



(a) Downlink



(b) Uplink

Figure 6.2: Comparison of surcharge values from all three experiments

lecting the most 'inefficient' user. For this, we decided on the basis of the downlink direction (Fig. 6.2a), as the uplink does not show a clear tendency regarding a single, worst candidate (Fig. 6.2b). The selected device is the one with the highest distance to the AP, having the largest surcharge value in uplink. After this single handover, i.e., 47 active VoIP nodes in total, the packet loss is below 5 percent in more than 90 percent of the evaluation intervals for downlink and uplink direction, respectively (Tables 6.1, 6.2). Thus, QoS constraints as defined in Section 6.4.5 are met for all 47 nodes due to a single handover following the inefficiency metric.

Now, let us consider the effect of a single replacement, i.e., the most inefficient node is triggered to perform a handover from WLAN to WWAN, while the WLAN network accommodates another VoIP node (with a distance of 10 meters to the AP). Figure 6.2 shows the surcharge values in up- as well as downlink direction for all three experiments.

The values increase with larger distances between STAs and AP. Also the SNR-values are tightly effected by the distance between a STA and its AP. As shown in Fig. B.3 in Appendix B.4, SNR monotonically decreases with the distance. Thus, under the assumptions of this simulation study, SNR- and inefficiency-based decisions select the same devices for a handover, as the probability for lower SNR values and thus lower link data rates and higher number of retransmissions increases with the distance. Note that all these impacts are now unified in the surcharge part of the inefficiency metric.

Not surprisingly, the 'max. #nodes' experiment results in highest surcharge values for all nodes, while the single replacement experiment leads to a significant reduction: the surcharge values drop by around 2.3 to 3.9 percent (downlink) and 2.9 to 5.9 percent (uplink). Lowest surcharge values are gained with the 'reduced set' experiment, where the most inefficient node was selected as handover candidate. There, the surcharge values of all other remaining nodes drop by 3.6 to 7.9 percent in the downlink and 3.5 to 12 percent in the uplink compared to "max. #nodes" results.

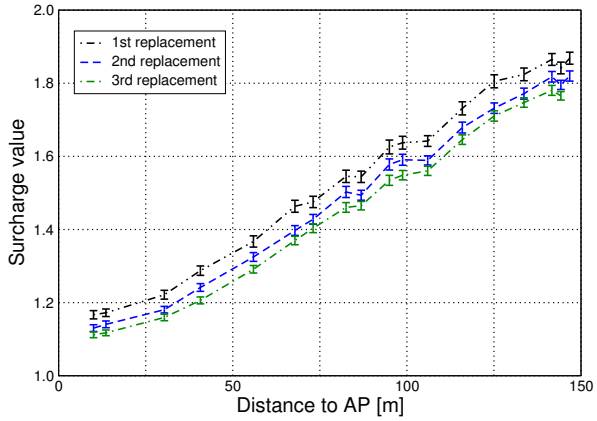
It attracts attention that surcharge values are higher for the up- than for the downlink direction. This stems from the *asymmetric traffic* conditions: the AP has to serve 48 VoIP streams in the downlink, i.e., 48 times more traffic than each single VoIP node in the uplink. This asymmetric traffic distribution leads to a lower a collision probability for the AP. Beside other aspects, Cai et al. investigated this effect analytically in their work [158]. The discrepancy between up- and downlink amplifies here, since the collision level has also an impact on the rate adaptation scheme.

The positive impact of further replacements is displayed in Fig 6.3, again for up- as well as downlink. While the second replacement leads again to a relatively large decrease, no significant improvement was gained by the third replacement (i.e., confidence intervals of the second and third replacements overlap at several distances).

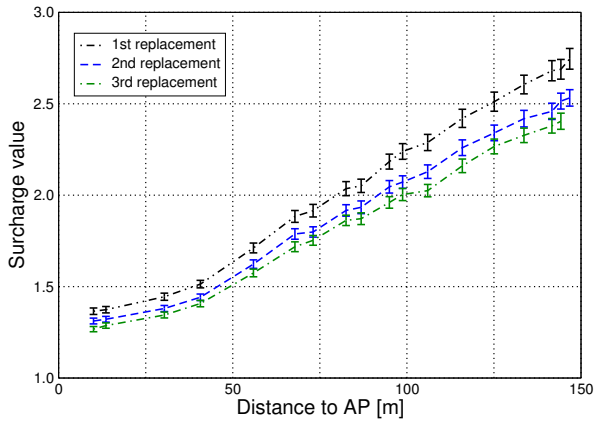
Lastly, we consider the impact of replacements on users' QoS. While the first replacement does not improve the application losses greatly for up -and downlink, it is the second replacement that avoids a violation of QoS constraints. From Tables 6.1 and 6.2, we can observe that less than 5 percent losses occur in 90 percent of the intervals for up- as well as downlink direction. Finally, the third replacement brings users' QoS up to level of the reduced-set experiment, which means that we gain comparable quality although there are 48 instead of 47 nodes. Interestingly, there are only small differences in QoS values between the second and the third replacement. This is directly in line with the surcharge results, where confidence intervals overlap such that there's no significant difference at certain points anymore. From the replacement study we observe that a non-significant impact of a replacement on the surcharge also implies only marginal differences in users' QoS in case of VoIP traffic.

6.4.7 Summary

Our proof-of-concept simulations document that the selected metric is suitable to select most 'inefficient' users in a scenario with *homogeneous* traffic patterns. Comparing it to SNR measures, it leads to the same decisions in such setups. The results also show the improvements for users remaining in the WLAN access cell, after performing a handover



(a) Downlink



(b) Uplink

Figure 6.3: Surcharge values after one, two, and three replacements

of the most inefficient candidate. Further, we studied the impact of our scheme in case of multiple handovers, where 'inefficient' WLAN users were replaced by suitable candidates from other heterogeneous access networks. Motivated by these improvements, we take the inefficiency metric as a basis for a more generalized onloading decision scheme that is presented in the subsequent sections.

6.5 From the Inefficiency to an Onloading Decision Scheme

The inefficiency metric has been designed to select 'bad' STAs by analyzing the actual appearing airtime of each. Although the inefficiency allows statements how efficient each STA uses the occupied airtime, it does not relate it to the overall load that a STA imposes on the channel. In other words, the metric does not rank onloading candidates

according to their load level. As a result, the inefficiency metric itself can be barely a 'stand-alone' approach as the statement about the efficiency of a STA can deploy its full potential only in conjunction with a second aspect reflecting the airtime for each device.

Let us consider a simple example to illustrate this point. Assume, we have a hotspot with four WLAN STAs, the first two driving high-quality video traffic, while the other two are running VoIP with the typical low throughput. Further, let us assume that one VoIP as well as one Video STA are wasting WLAN resources such that they show on average similar values in the inefficiency metric. In this case, the inefficiency does not show a clear ranking and thus does not identify a dominating handover candidate. A priori, it is even not clear which candidate should be favored as this also depends on the specific situation in the other technology, i.e., whether the WWAN rather favors a high-bandwidth consuming video or a low-rate VoIP flow.

6.5.1 Cost-Function Approach for Onloading Handover Decisions

To enable a flexibility for network operators at that stage, we combine the load measure and the inefficiency metric by a simple mathematical construct that is tunable regarding selectable operational points. We select the *simple additive weighting (SAW)* approach which is one of the most prominent mathematical approaches for "multiple attribute decision making" schemes for network selection [60]. Applying SAW for the two selected measures, the inefficiency and the airtime, leads to the cost function given below. Our cost function of the WLAN access cell evaluates the load together with the inefficiency of occupied resources for each device m,

$$c_{\mathrm{WLAN}}(m) = \omega_1 \frac{t_a(m)}{\Delta t} + \omega_2 \frac{D(m)}{D_{\max}(m)}, \text{ with } \omega_1 + \omega_2 = 1. \tag{6.9}$$

While D represents the inefficiency metric that evaluates the resource utilization on behalf of each traffic stream, the ratio $t_a/\Delta t$ consists of the airtime on the channel in relation to measurement interval Δt. Again, the airtime t_a represents the amount of time that the wireless medium has been occupied (or reserved, in case of inter-frame spaces and NAV settings) for the transmission of all packets K to and from device m during Δt:

$$t_a(m) = \sum_{j=1}^{K} t_{a_j}(m) = \sum_{j=1}^{K} \sum_{i=1}^{\text{trials}} t_{\mathrm{IFS}} + t_d(R_{i,j}) + t_{\mathrm{ack}}. \tag{6.10}$$

This includes the whole transmission sequence consisting of the inter-frame spaces DIFS or AIFS and SIFS (being included in t_{IFS}), the duration t_d of the complete data frame 'on air', where the data part is encoded with a certain modulation scheme $R_{i,j}$, and the acknowledgment t_{ack}. The number of trials represents the transmission attempts that have been required for the delivery of the MSDU.

6.5.2 Accounting for Link and Traffic Asymmetries

In the following, we present how we normalize and weight the metrics of the cost function. WLANs are usually faced with asymmetric links, i.e., the conditions for transmission in

up- and downlink direction are different. To accommodate this behavior resulting out of link and traffic asymmetries, we adjust the inefficiency values calculated for each up- and downlink, D_{up} and D_{down}, by weighting them with the relation of the airtime for each direction:

$$D = \frac{t_{\text{a}_{\text{up}}}}{t_a} D_{\text{up}} + \frac{t_{\text{a}_{\text{down}}}}{t_a} D_{\text{down}} \text{ , whereby } t_a = t_{\text{a}_{\text{up}}} + t_{\text{a}_{\text{down}}}. \tag{6.11}$$

We argue that the weighting of the inefficiency values by their airtime share in up- and downlink direction covers all reasons of the asymmetries which we discuss briefly in the following. One reason for asymmetries may be the multi-path propagation property of wireless channels, i.e., emitted wireless signals may have another dominant path for one direction than for the other one. On MAC level, asymmetries further appear as the WLAN AP as the central coordinator has to serve all of its associated STAs, thus accessing the wireless channel more frequently leading to a different collision probability for a traffic flow up- and downlink. In addition, the traffic sources further lead to different resource consumption in up- and downlink. For example, a TCP stream transports data packets up to a payload size of 1500 Bytes, while in the opposite direction short TCP-level acknowledgments are transmitted. Note that all these aspects are reflected by the airtime shares.

6.6 Performance Evaluation of the Decision Scheme

This section shows the gains of our approach further highlighting the impact of different cost-function weights on the operational state of the WLAN hotspot. More specifically, we tackle the following open issues regarding our onloading decision scheme: first, we demonstrate that, using our inefficiency approach, the maximum, offloadable number of traffic streams significantly improves compared to RSSI and random decisions. Second, we consider mixes of realtime and elastic traffic, i.e., VoIP and FTP streams, rather than homogeneous traffic alone. Third, as different decision schemes lead to different operational points of the WLAN network, our results allow operators to fine-tune the offloaded traffic mix such that it maximizes the utilization of WLANs.

6.6.1 Comparative Schemes

We firstly describe the approaches for WLAN, with which we compared our novel solution. We refer to them as 'comparative schemes'. Afterwards, we introduce selected flavors of our 'cost-function approach', which has been presented already above in Sec. 6.5.

RSSI For vertical handover decisions, the most common approach conducts a handover once the RSSI of a STA undergoes a certain threshold [59]. Accordingly, the first selection scheme relies on RSSI measurements in WLAN: For each device, RSSI-values are collected on the receiver side(s) for each successfully received data frame and averaged over Δt. STAs with the lowest RSSI-values are selected as handover candidates.

Random Selection The most simple approach offloads traffic to WLAN once the connectivity is available. To capture this somewhat simple behavior also for the backward onloading handover direction, we apply a decision scheme that selects STAs randomly from the 802.11 cell in a uniformly distributed fashion. In other words, each STA in WLAN has the same probability of being selected for a handover to WWAN.

6.6.2 Two Selected Flavors of the Cost-Function Schemes

Extending our work regarding the selection of inefficient WLAN users, this section considers two flavors of the cost-function approach, denoted as 'inefficiency' and 'equal weight (EW)' decisions, which differ regarding the weights of the two cost function metrics, the occupied airtime and the inefficiency value of a WLAN STA. The first flavor only takes into account the inefficiency of the wireless transmissions (i.e., $\omega_1 = 0$, $\omega_2 = 1$ in Eq. 6.9). In order to penalize inefficient traffic streams evoking a high channel load, 'EW' will consider the impact of a mix of inefficiency and wireless channel load, measured by the occupied airtime, by setting $\omega_1 = 0.5$, $\omega_2 = 0.5$ in Eq. 6.9.

6.6.3 Methodology

We compare the performance of the four selected decision schemes regarding the number of VoIP flows and the volume of data traffic that can be accommodated by a WLAN cell thus unloading WWAN. To trade off both traffic types against each other, we additionally consider the question how much FTP traffic one can transport at the costs of a reduced number of VoIP clients. For this, we applied a two-stage process: we first determined the capacity of the WLAN cell in terms of VoIP users which can be simultaneously served without a violation of QoS constraints. In the second stage, we considered VoIP and FTP traffic mixes in the WLAN cell. On this basis, we finally compared the operational point of each traffic mix resulting from each selection scheme with the capacity from the pure VoIP scenario.

6.6.4 Simulation Model

In our simulations, we modeled the WLAN hotspot of our scenario given in Sec. 6.1. Again, the WLAN hotspot is represented by an 802.11g AP that is 11e-capable by providing EDCA functionality. The applied EDCA simulation model is described in detail in Appendix B.1. In the simulation scenario, WLAN devices are distributed uniform randomly over the squared area of interest. There, the AP is located at the corner of the considered environment, such that no hidden nodes appear. All WLAN devices apply 802.11g ERP-OFDM—with link data rates from 6 up to 54 Mbps, whereby we model a perfect rate adaptation regarding the RSSI. The 802.11e/g parameters were chosen according to [91], leading to the backoff parameter given in Table 6.3[4]. We used AC_VO for VoIP traffic, while AC_BK backoff parameters were applied for FTP traffic. The delay between the AP of the WLAN access cell and the wired nodes was set to 100 ms.

[4]TXOPLimits were set to zero so that a single transmission per medium access attempt is performed

Table 6.3: Backoff parameter set

AC	CW_{min}	CW_{max}	AIFSN
AC_BK	15	1023	7
AC_BE	15	1023	3
AC_VI	7	15	2
AC_VO	3	7	2

To take into account that radio signals are not only affected by path loss but also by multipath propagation, we applied our ns-2 simulation suite including a log-distance path loss, a Ricean fading, as well as an SINR model as detailed in Appendix B.

Further, we assume for this simulation study to have exact knowledge about the parameter values thus not modeling any 802.11k signaling. At the end of this chapter, in the final remarks in Sec. 6.7, we discuss practical considerations for measuring and signaling the parameters for airtime and inefficiency. There, we argue that by means of 802.11k we can obtain precise information regarding the selected measures.

Modeling Off- and Onloading Decisions We now describe the process by which we decide when to admit a device in WLAN (offloading) and when to conduct a handover back to WWAN (onloading). Basically, we apply the following algorithm in our simulations: we evaluate at each interval Δt_{HO} whether the QoS values of the served STAs in WLAN have been violated. If no violation has occurred, we accommodate an additional user in the WLAN hotspot by selecting him in a random-uniform fashion from all STAs being within WLAN coverage as well as being connected to WWAN. In case that QoS limits have been violated for at least one STA, we calculate the respective performance metric. Then, we handover the STA with the worst metric value from WLAN. To account for the situation that QoS limits of a STA are violated because of a bad channel instead of an overload situation, we introduced a QoS penalty. That is, if the QoS for a certain STA has been violated continuously for more than five Δt_{HO}, it is selected for a handover, too. Throughout this work, we apply a handover interval Δt_{ho} of one second.

Traffic Model and QoS Constraints In this chapter, we consider VoIP traffic as well as data transfers via FTP. Data traffic was generated by FTP clients, which either down- or uploaded a large file of infinite size via TCP. TCP/IP segments had a size of 1500 Bytes and the TCP-SACK option was used. For the transmission of TCP/IP data segments in WLAN, we set a hard minimum MAC-level goodput of 128 kbps obtained over a period of the last four seconds. By this, we aim to enable a transmission on average of at least one TCP data packet (and accompanying TCP ACKs) per WLAN beacon interval.

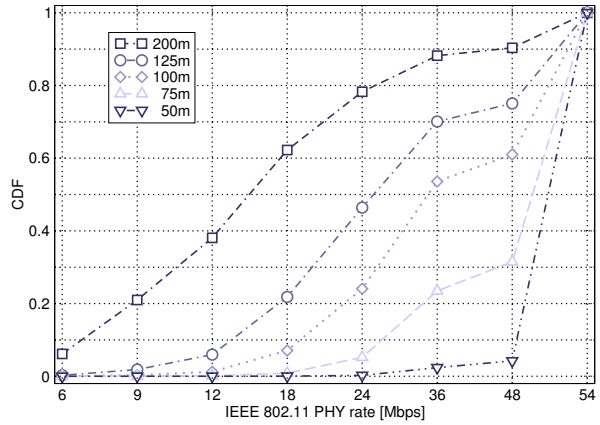


Figure 6.4: Distribution of 802.11g PHY rates for different edge lengths

For VoIP, we used the same model, the parameterization, and similar QoS limits as described in Sec. 6.4.5, i.e., an exponential ON/OFF model generating packets according to the G.711 codec (160 Byte audio packets each 20 ms during ON periods). Again, we consider the QoS limit for VoIP in terms of losses, consisting of lost and late packets. If five or more percent of the VoIP packets are lost, the quality is assumed to be lousy. To account for the ON/OFF patterns of the VoIP traffic, we calculate the VoIP QoS measure as well as onloading decision metrics over the last ten handover intervals.

Scenarios We investigated three scenarios with different traffic mixes: pure VoIP, VoIP with FTP downloads, and VoIP with FTP uploads. In each setting, 200 devices are distributed of the area of interest. For this, we draw device positions from uniform random distributions for the x and y coordinates. For all three traffic scenarios, we vary the size of the considered quadratic area with edge lengths of 50, 75, 100, 125, and 200 meters. Thereby, we scale the x and y coordinates of the devices by the edge length, while the AP resides at a corner of this area. In the pure VoIP mix, all devices run a VoIP call. For the other traffic mixes, we randomly select 150 devices to carry VoIP and the remaining 50 to run FTP traffic. Note that larger edge lengths lead to greater sizes of the areas such that the link data rates applied for 802.11 transmissions are more and more reduced. The cumulative distribution function (CDF) of the link data rates for the different edge lengths is shown in Fig. 6.4.

6.6.5 Performance Metrics

Let us shortly discuss the selected performance metrics. To enable a comparison of the different schemes, we do not only consider the number of clients and the MAC-level goodput but do also introduce a measure relating both.

Number of Clients For an understanding about the number of supported clients from different traffic types, we consider the number of VoIP and FTP clients, which reside within the WLAN hotspot having satisfied QoS constraints.

MAC-Level Goodput To observe the load situation in the WLAN cell, we consider the aggregated MAC level goodput of FTP data packets at the end of each Δt_{ho} calculated over the last second. This gives an indication how much elastic traffic is transmitted successfully at layer 2.

Goodput/VoIP-Reduction Ratio (GVR) To allow a comparison of the decision schemes in the scenarios with traffic mixes of VoIP with elastic traffic such as FTP, we introduce the goodput/VoIP-reduction metric. It is a revenue/cost ratio metric, where the costs are the number of VoIP calls that were removed from the WLAN cell. In contrast, the revenue is the aggregated MAC goodput of the FTP traffic ($G_{\text{FTP, MAC}}$) that is accommodated by the WLAN cell instead.

In other words, since we consider the capacity of a WLAN network in terms of VoIP calls, we define a measure that gives us the revenue for the case that we replace certain VoIPs by FTP clients in the WLAN network:

$$\text{GVR} = \frac{G_{\text{FTP, MAC}}}{N_{\text{VoIP}_{\text{max}}} - N_{\text{VoIP}_{\text{cur}}}}. \tag{6.12}$$

There, $N_{\text{VoIP}_{\text{max}}}$ is the capacity of the network in terms of VoIP users (from the VoIP only case), while $N_{\text{VoIP}_{\text{cur}}}$ is the current number of served VoIP clients by WLAN in the scenarios with the traffic mixes.

6.6.6 Evaluation Procedures

The results of this study base on extensive simulations with 1200 runs, each one lasting about a day. For the evaluation, we conducted *independent replications* [153] with five repetitions. After deleting the transient phase, we evaluated the data regarding the number of VoIP and FTP STAs as well as the aggregated MAC goodput with a 90 percent confidence level.

For the GVR metric, where the denominator is not just a constant, a proper computation of confidence intervals suffers from the fact that ratios of (sample) means do not obey a normal distribution anymore [176]. As a solution, we applied Fieller's method [177]. For a short discussion, we follow the illustrative survey about Fieller's method as given by Franz [176]. Basically, the method utilizes the property of the difference between two normally distributed random variables x and y, which obeys a normal distribution again. As result, with a ratio $\rho = \hat{y}/\hat{x}$, where $\hat{x}$ and $\hat{y}$ are the mean values of x and y, also the difference $\hat{y} - \hat{x}\rho$ is normally distributed. Normalizing this difference with the joint standard deviation of $\hat{y} - \hat{x}\rho$ to a standard normal random variable leads to

$$t_{1-\alpha/2} = \frac{\hat{y} - \hat{x}\rho}{\sqrt{\rho^2\sigma_x^2 + \sigma_y^2 + 2\rho\sigma_{x,y}}}, \tag{6.13}$$

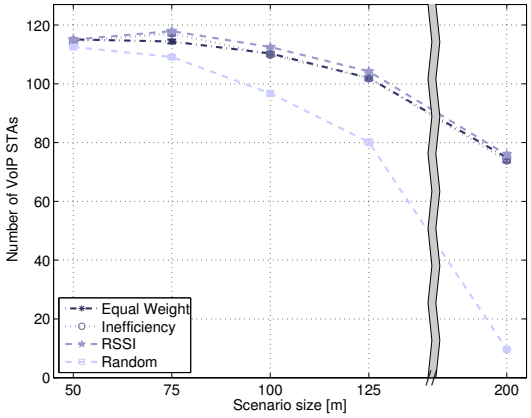


Figure 6.5: VoIP capacity with the different decision schemes

where $\sigma^2_{\hat{x}}$ and $\sigma^2_{\hat{y}}$ are the variances of x and y, $\sigma_{x,y}$ is the covariance of x and y, and $t_{1-\alpha/2}$ represents the selected $(1-\alpha/2)$ quantile of the t-distribution [176, 178].

If $\hat{x}^2/\sigma^2_{\hat{x}} > t^2_{1-\alpha/2}$, Fieller was able to determine the confidence interval bounds $\rho_{1/2}$, by solving Eq. 6.13 towards ρ [176]:

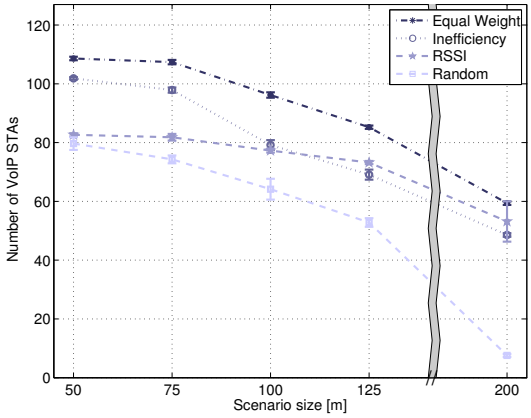
$$\rho_{1/2} = \frac{\hat{x}\hat{y} - t^2_{1-\alpha/2}\sigma_{x,y}}{\hat{x}^2 - t^2_{1-\alpha/2}\sigma^2_{\hat{x}}} \pm \frac{\sqrt{(\hat{x}\hat{y} - t^2_{1-\alpha/2}\sigma_{x,y})^2 - (\hat{x}^2 - t^2_{1-\alpha/2}\sigma^2_{\hat{x}})(\hat{y}^2 - t^2_{1-\alpha/2}\sigma^2_{\hat{y}})}}{\hat{x}^2 - t^2_{1-\alpha/2}\sigma^2_{\hat{x}}} \tag{6.14}$$

Within this thesis, our results always fulfilled the condition of $\hat{x}^2/\sigma^2_{\hat{x}} > t^2_{1-\alpha/2}$. Regarding the analysis of other cases of Fieller's method, where the denominator may include zero, the reader is referred to von Luxburg and Franz [178].

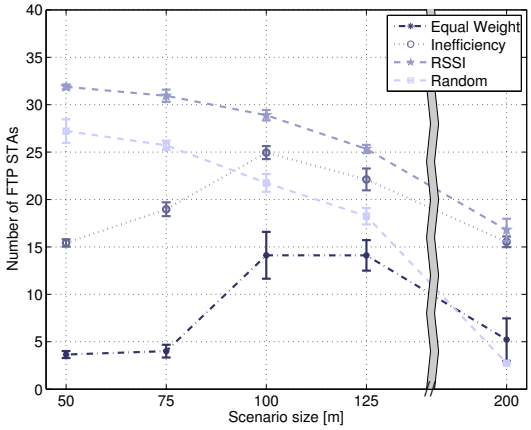
6.6.7 Results for Pure VoIP and Traffic Mixes

Fig. 6.5 shows the results for the first simulation setup with pure VoIP traffic. With increasing sizes of the area, the number of VoIP clients reduces for all decision schemes. For RSSI and both cost-function schemes, the VoIP capacity behaves similar. This is a result of the homogeneous traffic, for which only the surcharge value of the inefficiency metric may have an impact, as the overhead factor stays constant for all traffic flows. This confirms our results from Sec. 6.4.6 where the surcharge value increases with the distance between AP and STA due to the increasing probability for low rate transmissions and higher number of retransmissions. Only the VoIP capacity with the random selection drops significantly faster, which is a result of the technology-agnostic decisions.

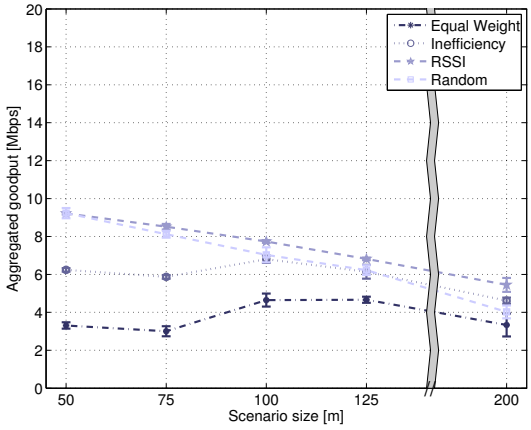
The range of simultaneous VoIP calls for RSSI and cost-function decisions corresponds with results of previous work [179] that showed 105 VoIP calls for pure 802.11g



(a) VoIP STAs

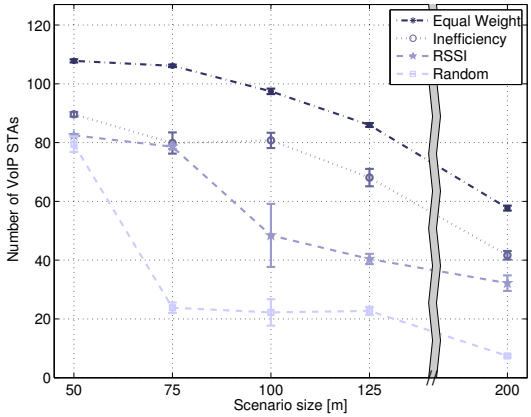


(b) FTP STAs

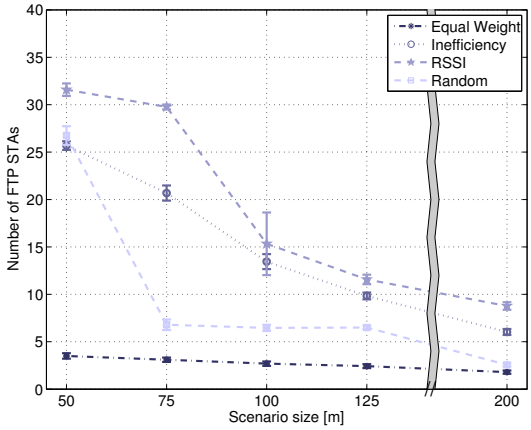


(c) Aggregated MAC goodput for FTP STAs

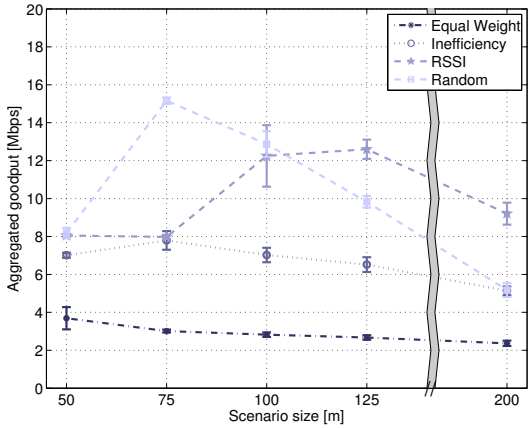
Figure 6.6: FTP downloads together with VoIP traffic



(a) VoIP STAs



(b) FTP STAs



(c) Aggregated MAC goodput for FTP STAs

Figure 6.7: FTP uploads together with VoIP traffic

ERP-OFDM scenarios. We gain a little higher capacity values for scenario sizes up to 100 meters, as we apply the 802.11e EDCA MAC protocol, which has smaller $CW_{\min}$ and $CW_{\max}$ values of three and seven. Thus the AP stays on average less time in the (post-)backoff process enabling some more calls to be served.

VoIP-FTP Traffic Mixes

Results for VoIP plus FTP download and upload traffic are shown in Fig. 6.6 and 6.7. First, let us consider the number of VoIP STAs in Fig. 6.6a and 6.7a. In both, the EW decision scheme comes up with the highest number of VoIP users, followed by inefficiency, RSSI and random decision schemes.

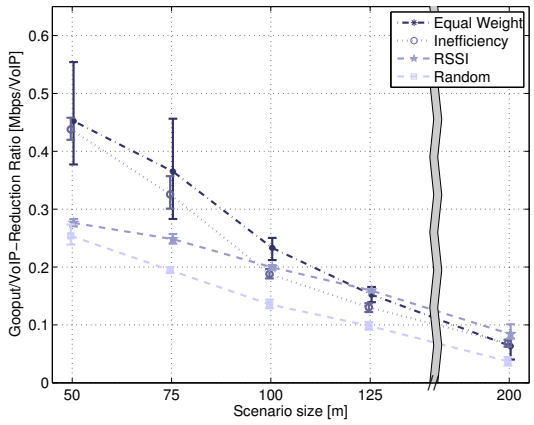
Consequently, the number of FTP users, shown in Figure 6.6b and 6.7b, are the smallest for the EW scheme. With the RSSI scheme, the highest number of FTP users are gained, while random and inefficiency decisions range in between.

To complete the overall picture, Figs 6.6c and 6.7c show the aggregated MAC level goodput for all accommodated FTP STAs. For FTP downloads, the aggregated MAC goodput curves follow the shape and the relations of the number of FTP STAs pretty closely. In contrast, for the FTP uploads, random and RSSI decisions have high peaks in the goodput curves at 75 and 100 to 125 meters, which corresponds with the great reduction regarding the number of VoIP STAs at these points.

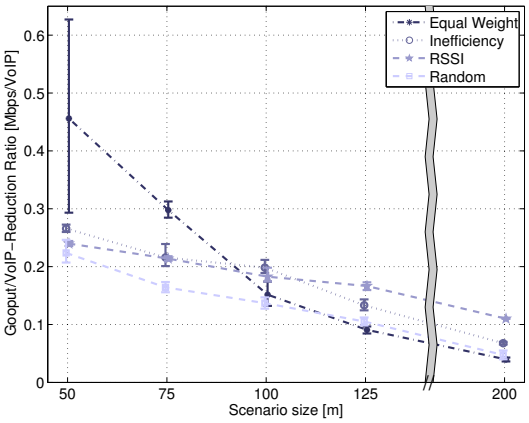
Finally, we can conclude that the differences in the aggregated FTP MAC goodput lead to different operational points of the network. As one can see from Fig. 6.6a and 6.7a, also the VoIP curves are affected. The difference in aggregated FTP goodput is a result of the MAC scheme which ensures fairness on a per station and traffic class basis. In hotspots, the AP serves all *VoIP and FTP STAs*. Thus, the AP has to contend with the uplink traffic, which results in higher queuing delays and drops for FTP data packets in the downlink. As this lowers the aggregated FTP goodput, it leaves more capacity for VoIPs. Contrary, FTP uploads have higher aggregated goodputs, leading to a smaller number of VoIPs as their QoS constraints are violated earlier such that handovers are conducted earlier. In the 802.11 area, the phenomenon of higher delays and drops in the downlink is known as *asymmetry problem* and effects have been identified in the literature for VoIP [180] (also compare Sec. 6.4.6) and TCP traffic [181].

Goodput/VoIP-Reduction Ratios

Lastly, we consider GVR for FTP up- and downloads in Fig. 6.8. Note that the relatively large confidence intervals are a result of the denominator ΔN_{VoIP} which is the difference between $N_{\text{VoIP}_{\max}}$ and $N_{\text{VoIP}_{\text{cur}}}$. Although the numbers of VoIP STAs itself have small variances and confidence intervals (cf. Fig. 6.5, 6.6a, and 6.7a), these small variances lead to a higher impact on GVR as they reside in the denominator. In other words, the high confidence intervals for EW and inefficiency schemes stem from small ΔN_{VoIP}. In scenarios where ΔN_{VoIP} becomes larger, the confidence intervals for GVR behave similar to the other schemes.



(a) Mix of FTP downloads with VoIP traffic



(b) Mix of FTP uploads with VoIP traffic

Figure 6.8: Goodput/VoIP-reduction ratios

Now, for the FTP downloads, both equal weight and inefficiency decisions outperform RSSI and random for the first two scenario sizes. While equal weight is also better at 100 meters, inefficiency and RSSI do not differ significantly anymore. For the remaining scenarios, equal weight, inefficiency, and RSSI perform similar. Despite the smallest and largest scenario, random decisions lead to the worst results, which is not surprising as the number of VoIP clients was stronger reduced than with the other schemes.

For FTP uploads, EW outperforms again for the first two scenarios, while the inefficiency strategy is slightly better than RSSI for the smallest scenario. In contrast, RSSI decisions are the best for large-sized scenarios, where all other schemes behave similar but worse than RSSI.

Overall, it is surprising that the well accepted RSSI-metric for handover decisions is better only for FTP upload traffic in very large, rather unlikely scenarios. For more realistic cases, the results show improvements for our decision schemes.

6.6.8 Conclusions

We analyzed the performance gains of our novel scheme for selecting WLAN STAs to be onloaded back to WWAN, thus in turn allowing to maximize the amount of traffic being offloaded to 802.11 hotspots. In our performance evaluation, we compared the gains of our approach with the classical RSSI as well as random decisions. Overall, our scheme outperforms all others for the dense settings in which 802.11g STAs transmit at medium to high PHY rates. In sparse settings with FTP downloads, where the AP covers a large area, all decision schemes performs similar. Only with FTP uploads in sparse settings, RSSI decisions are slightly better, being essentially an effect of traffic asymmetries. Additionally, our results show a tendency for operational points that is of interest for mobile operators: if only few flows with elastic data traffic should be served, the equal weight strategy is the right choice, whereby the inefficiency scheme may be used for more data flows but fewer VoIPs.

6.7 Final Remarks: Practical Concept for Obtaining Airtime and Inefficiency

Measuring and calculating the two main parts of our cost function, the airtime and the inefficiency, requires to have access to dedicated pieces of information from MAC level. In the following we shortly describe which parameters we need to extract and signal in practical networks at certain points. We emphasize that we focus just on simple measures being readily available on today's hardware, which are among others, the size of the data part of a frame, the number of transmission attempts, the link data rate, and information about the successful transmission of a frame [182].

The computation of the cost function values given in Eq. 6.9 bases on the airtime t_a from Eq. 6.3 and the smallest possible transmission duration Δt_{opt} from Eq. 6.5. The key issue is how to obtain these values on AP side, for all traffic of a given STA in up- and downlink direction.

6.7.1 Downlink Transmissions by the AP

Let us start the discussion from the view point of an AP, which is the central coordinator of a WLAN hotspot. In the downlink direction, the AP serves all associated STAs by transmitting traffic for each in the downlink direction. In other words, the AP has a complete knowledge about all the parameters in the downlink. As a result, the AP can easily conduct statistics per STA or per flow basis for Δt_{opt} by the "Transmit Stream/Category Measurement" detailed in Sec. 3.7. These statistics include in our case the frame size, from which we can deduce Δt_{opt} as all other parameters in Eq. 6.5

are known for a given WLAN PHY. Further, regarding the airtime t_a per STA, we assume that the AP is capable to trace the amount of time for its downlink transmissions to a given STA. For a signaling of this airtime, we make use of one of the reserved fields of the "Transmit Stream/Category Measurement" for proprietary measures.

6.7.2 Uplink Traffic of STAs

For uplink directed traffic from the STAs, we can derive Δt_{opt} from the size of received uplink transmissions. However, the AP may have incomplete statistics regarding the conducted number of retransmissions and the link data rates used for their transmission, which strongly affects the airtime measure t_a of a STA in the uplink. Accordingly, we discuss in the following, how we obtain the airtime of a STA for its uplink transmissions.

While the AP can determine all the related parameters for *successfully* received frames, it cannot extract these pieces of information for data frames which have not been decoded correctly. If even all subsequent transmission attempts of a data frame fail, i.e., the transmitting STA accordingly drops the frame after the last unsuccessful attempt, the AP is capable of detecting the missing frame by a gap in the sequence number space being revealed with the reception of the next data frame. In such a case, the AP knows at least that the STA has conducted its maximum number of retries without properly transmitting the uplink frame. Overall, the AP can deduce the number of frames N_{losses}, for which the maximum number of transmission attempts have been unsuccessful.

Further, let us consider the case where the STA has to retransmit a frame but is finally successful. For this case, it is important to note that all frames which are retransmitted in 802.11 networks have been specifically marked by a retransmission flag. If such a frame is correctly received by the AP, it knows that at least two transmission attempts have been conducted by the STA, which 'lifts the fog' about erroneous transmissions to a certain extend. An issue still remains if a STA has to transmit the same data frame more than twice as a result of erroneous transmissions. Note that the cases of single and multiple retransmissions cannot be easily distinguished by the AP. To deal with these situations, we have three options:

1. Operating with incomplete knowledge, i.e., neglecting more than two transmission attempts for a concerned frame,
2. Collecting measurements from other STAs about uplink transmissions they have been overhearing, and
3. Obtaining statistics from the involved STA about its uplink transmissions.

Incomplete Knowledge

The first option is the most simple one, however neglecting the effort of more than two transmission attempts for the transport of a data frame. Although some related work argues that STAs should trigger timely handovers themselves after just three consecutive transmission attempts [126], we note that this does not hold in general for all types of

WLAN STAs as these algorithms are not standardized and thus highly vendor specific. As a result, the impact of the first option operating with incomplete knowledge depends on the mix of present STAs from different vendors in the hotspot thus making a proper evaluation about their impact on the (unnoticeable) number of transmission attempts impossible. Nevertheless, this option may be used if no information from STA side is available. As STAs with high number of retransmissions tend to lower their link data rates down to the most robust MCS, there is a high probability that such STAs will also be selected as handover candidates with the first option.

Measurements from Overhearing STAs

Uplink transmissions of our STA that have not been received properly by the AP, may have been received nevertheless by other STAs. The proper reception of frames not being destined to itself is denoted as 'overhearing'. This could be utilized by the second option. Information of frames being overheard by other STAs may conceptually help in our context. Although certain approaches in the 802.11 world exist to support the collection of distributed, measured information (e.g., the 802.11k "Frame Report" [92]), this does not help much as the relevant pieces of information, i.e., number of retransmissions and applied link data rates for overheard frames with the retry flag set, are not supported at all. Further, such a distributed measurement among many WLAN STAs may lead to significant signaling overheads the more STAs are involved in this process.

Evaluation of Statistics from a STA

Finally in the last case, only the involved STA maintains statistics for its uplink transmissions, which have to be signaled to the AP. This is inline with the context of the 802.11k amendment, detailed in Sec. 3.7.1. Recall that 802.11k introduced so called "STA Statistics". With 802.11k, the AP is able to trigger measurements on STA side for a given measurement duration. After this phase, the STA sends the measured values back to AP by its response frame.

Among others, these STA Statistics include cumulative counters, of which three are important for us: "dot11RetryCount", "dot11MultipleRetryCount", and "dot11ACKFailureCount". The first gives the number of data frames for which "one or more retransmissions" were needed, while finally an ACK was received for each. In contrast, the second specifies the number of data frames for which "more than one retransmission" was required until obtaining an ACK. Finally, the third gives the total number of transmission attempts (including non-successful retransmissions) for which an outstanding ACK was not received [6].

Obtaining Retransmission Statistics From the first two counters we can easily derive the exact number of frames for which just one retransmission was required, denoted as RTX_1 in the following and given by

$$\mathrm{RTX}_1 = \mathrm{dot11RetryCount} - \mathrm{dot11MultipleRetryCount}. \tag{6.15}$$

While we know the number of successful frames that required more than one retransmission (dot11MultipleRetryCount), we still need an approximation regarding the number of transmission attempts conducted for each frame. We extract an average value from all aspects available at AP. There, on the AP side, we know the total number of successfully transmitted frames M_{RTX} of the STA that required one or more retries (from the retry flag in the frame control field of the uplink frames). Thus, we can infer the mean number of multiple consecutive retransmission attempts of the STA by

$$\mu_{\mathrm{multiRTX}} = \frac{\mathrm{dot11ACKFailureCount} - \mathrm{Retries}_{\mathrm{max}} N_{\mathrm{losses}} - \mathrm{RTX}_1}{M_{\mathrm{RTX}} - \mathrm{RTX}_1}, \tag{6.16}$$

whereby $\mathrm{Retries}_{\mathrm{max}}$ is the maximum number of retransmissions after which a STA drops the frame. We set $\mathrm{Retries}_{\mathrm{max}}$ to the retransmission limit(s) as given by the 802.11 standard [6]. We are aware that the standard gives only default values. The maximum retry limit may indeed be dependent on the STA configuration and the applied link data rate algorithm. However note that if $\mathrm{Retries}_{\mathrm{max}}$ for a STA takes other values than the default, this has a direct impact on the mean value in Eq. 6.16. In case of higher $\mathrm{Retries}_{\mathrm{max}}$ the mean values lowers; for smaller $\mathrm{Retries}_{\mathrm{max}}$, μ_{multiRTX} increases, while the maximum number of retransmissions is already attributed to the lost packets N_{losses}. Thus, we do not neglect retries on average.

To summarize, we are able to determine on the basis of 802.11k counters the exact number of frames for which exactly one (RTX_1) and the number for which multiple (dot11MultipleRetryCount) retransmissions were required. For the latter, we have been able to further deduce the mean number of retransmissions (μ_{multiRTX}).

Mapping of Retransmissions to Link Data Rates In a nutshell, the big picture regarding the number of required retransmissions becomes quite precise with 802.11k. Nevertheless, the applied link data rate for each unsuccessful transmission remains unknown for the AP. Furthermore, the 802.11k amendment does not standardize suitable measurements reflecting the link data rate selection of a STA. Conceptually, one could design own 802.11k counters on STA side, however it remains questionable whether all vendors will support this in their devices. Thus, we propose a simple heuristic for the 'link data rate to retransmission mapping' instead. It works as follows. The AP tracks link data rates and retransmission information for successful transmissions during the measurement period. For each received frame marked with the retry flag, we calculated the distance in the link data rate space between the data rate of the penultimate and the last successful transmission. This results overall in a list of retransmitted frames which we order according to the link data rate distance. Then, for the first dot11MultipleRetryCount entries with the largest distance in the link data rate space, we assume a number of (rounded) μ_{multiRTX} retransmissions all being conducted at the last successful rate. For other entries in our list, we assume single retransmissions at the last successful rate. Finally for lost packets, we assume that the maximum number of transmission attempts were conducted at the lowest possible link data rate. Note that with this 'link data rate to retransmission mapping' we penalize on average STAs with

lots of retransmissions as we assume that the last, usually lower link data rate has been applied for all unsuccessful transmission attempts of a frame, too.

Reducing Signaling Overheads Finally, to keep 802.11k signaling overheads low, we suggest to apply the third option, the evaluation of statistics on a per STA basis, not continuously for all associated WLAN device in the hotspot but on demand, i.e., for selected STAs showing a certain level of successfully received frames with the retry flag set. Recall that we aim at selecting the inefficient WLAN STAs. A significant number of retries usually appears either for STAs always transmitting at the lowest link data rates or for STAs sometimes drastically reducing their sending data rate, i.e, with large gaps in the link data rate space between two subsequent successful transmission, where the last has the retry flag set to one.

Chapter 7

Link Data Rate Estimation

Similar to other wireless technologies, a key issue for successful WLAN transmissions is the signal-to-interference and noise ratio (SINR). In reality, SINR fluctuates as a result of fading and interference, therefore modern wireless technologies usually incorporate schemes trying to adapt the transmission parameters to the SINR conditions. The IEEE 802.11 standard [6] specifies distinct sets of MCSs for each PHY specification (i.e., 802.11a/b/g/n). Determined by the applied PHY configuration, the transmitter is free to select a MCS on a per-packet basis for each transmission from its set. Thereby, different MCSs result in different raw bit rates for a transmission on PHY level. We refer to these bit rates as *link data rates* or *data rates* in short. Further, the policy for the selection of MCSs is referred to as the *rate adaptation (RA) scheme*.

While specifying different MCSs, the IEEE 802.11 standard itself does not give strategies and mechanisms for the RA. As a result, individual WLAN NICs of multiple vendors may apply different RA algorithms. Vendors compete in usage of proprietary solutions aimed at achieving a leading performance. Thus, an AP in a WLAN hotspot serving STAs with NICs from different vendors is confronted with a heterogeneity of behavior—without having any knowledge of the applied algorithms on the STA side.

Even a minor number of STAs transmitting at low-rates may harm the overall capacity of the WLAN cell. Medepalli et al. [183] showed that on the costs of three 54 Mbps VoIP users, roughly just a single low-rate 6 Mbps user can be served in their setting. However, no available mechanisms for an AP exist to enquire the applied RA scheme and its behavior from a WLAN NIC. Furthermore, it is questionable whether such an inquiry would be practically feasible as hundreds of different devices from various vendors may make a standardization as well as an implementation of proper semantics and syntax describing individual RA schemes quite complicated.

For WLAN hotspots, it would be highly beneficial to have knowledge about the RA behavior of all the WLAN cards involved, as it directly relates to the question how traffic from a given device impacts the hotspot load. In the context of the discussion regarding on- and offloading traffic from and to a WLAN hotspot, the RA behavior of devices appears in two different dimensions. First, it is relevant regarding admission decisions

for associating devices, whereby the knowledge about RA behavior could be a basis for decisions aiming to avoid devices in a WLAN hotspot which transmit only with low-rate MCSs. Second, also regarding present, active STAs in a WLAN hotspot, knowledge about their RA behavior helps analyzing a given situation, e.g., regarding the impact of an additional association or the long-term behavior of a STA.

This chapter presents a solution, denoted as data rate estimation (DARA), which allows a WLAN hotspot to estimate the rate selection behavior of a STA by an approach from the area of machine learning. More precisely, we observe a STA on short time scales in the order of just some beacon intervals and make an 'educated guess' regarding its RA behavior afterwards by machine learning methods. We define *RA behavior* as the probability mass function (pmf) of data rates for successful uplink transmissions of a given STA. Thereby, we assume that the traffic of the considered STA is stationary and has constant data packet sizes. Further, we assume that DARA does not have any knowledge of the internal operation of the RA scheme on STA side. To our best knowledge, this is the first solution offering such possibility.

Following our reference scenario with the architecture from Sec. 4.4 and 4.5, we focus in this chapter on an individual WLAN hotspot that applies our architectural framework given in Sec. 4.5.2. To verify our DARA approach, we consider the following three questions in this scenario:

1. What is the accuracy of DARA, when comparing its predictions with real values regarding the RA behavior for settings including various STA positions and different RA algorithms?

2. Combining DARA's estimates for associating devices with admission decisions on WLAN side, what are the gains that can be achieved in terms of additional devices to be served, compared to classical RSSI decisions?

3. For admission decisions, how often are DARA's decisions regarding the RA behavior correct?

The chapter is structured as follows. First, we present the refined scenario and assumptions in Sec. 7.1. After the related work in Sec. 7.2, Sec. 7.3 discusses the principle of DARA, before we present the machine learning model in Sec. 7.4. Sec. 7.5 describes the settings for DARA's evaluation regarding the first question, leading to the results in Sec 7.6. Then, Sec. 7.7 presents an application example for the second question, in which we utilize DARA for a selection of flows to be served by a WLAN hotspot. We demonstrate that using the information about RA instead of the usual RSSI allows us to admit more flows while assuring the same QoS level of the individual flows. In this context, we also analyze the fraction of correct decisions regarding our third question. Finally, Sec. 7.8 gives considerations for practical setups, while Sec. 7.9 concludes this chapter.

This work was published in a condensed version in [184], while the extended work [185] included details also given in this chapter regarding the approach, the parameter selection, the 'proof-of-concept' evaluation, and the application example consisting of an access technology selection scheme.

7.1 Refined Scenario and Assumptions

As mentioned above, we focus on an individual WLAN hotspot that applies DARA. Nevertheless, we assume that a number of other hotspots also operates on interfering WLAN channels. The number of these hotspots varies throughout this chapter and given below in each section separately. In addition, the area of the WLAN hotspots is covered by a WWAN with 3GPP 3G/4G technology.

Similar to the previous chapter, all end users are static and have multi-standard devices supporting both wireless accesses, WLAN as well as WWAN. Further, each active end device has an on-going traffic stream in up- and downlink direction, both either via WLAN or WWAN. In addition, we assume that the number of active end devices within the coverage area of the 802.11 hotspot exceeds the WLAN capacity in terms of accomplishable devices by far. Thus, we follow the assumption of the previous chapter: To handle high-load situations in WLAN, we presume that an admission control scheme on WLAN side regulates access for devices attempting to associate.

Further, we assume that a significant fraction of devices has stationary traffic in terms of packet inter-arrival times and uses constant packet sizes for transmissions. For this work, we study estimates for VoIP traffic as its strong QoS requirements make it interesting for a support of timely decisions of onloading handovers from our perspective. The number of active devices within the joint coverage area of WWAN and WLAN is given in each section below.

Regarding the architecture given in Sec. 4.4, we assume to make DARA's computations on the RMC. Thereby, we presume that the RMC is capable to conduct the required calculations in the order of some hundred milliseconds.[1] As a basis for the calculations on the RMC, the AP and the WLAN NIC of the end-user device conduct measurements of selected 802.11 parameters. Considerations for a practical usage regarding a signaling of these parameters are given in the final remarks at the end of this chapter.

7.2 Related Work

Comparisons of different, practical rate adaptation algorithms in the literature rely on schemes being implemented in the WLAN Linux driver *Madwifi* [186], which comes with the algorithms *Minstrel, Onoe, SampleRate,* and *AMRR*. Yin et al. [187] studied these algorithms in a wired setup, where wireless links were emulated by cables and adjustable attenuators. By changing the attenuation as well as adding interference patterns, the authors illustrated the strong differences between the operational points of the rate adaptation algorithms. Further, they identified that the Minstrel algorithm outperforms all others specifically in the presence of varying channels due to interference. Ancillotti et al. [188], among others, analyzed the impact of diverging rate adaptation in congested situations. By their indoor measurements, the authors showed that the aggregated throughput in a WLAN strongly varies among setups conducting either AMRR, SampleRate, or Onoe.

[1] In principle, the RMC may also rely on computations from cloud services to fulfill this assumption.

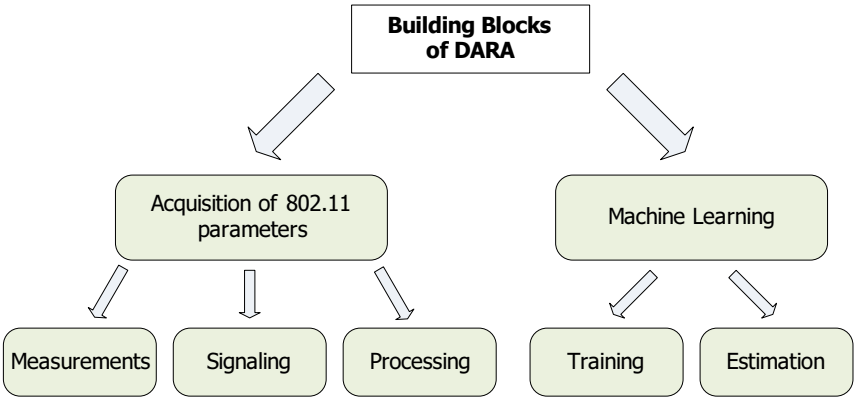


Figure 7.1: Basic building blocks of DARA

Relatively few work is available regarding the estimation of RA schemes. These solutions require a priori knowledge about the internal operation of the candidate RA scheme. Recently, Kim et al. [189] presented an analytical Markov chain model for WLANs including a prediction of applied data rates for the selected scheme Minstrel. Their estimation bases on the packet loss rates and remodels on this basis the internal behavior of the considered RA scheme. Further, Mirza et al. [190] presented a scheme that allows for a classification of RA schemes on the basis of their observed MAC parameters, which the authors denote as a "fingerprinting" of RA schemes. Their approach bases on specific, internal patterns with which changes of rates for each scheme appear. As the authors target a precise classification of each algorithm reactively on the basis of long observations, i.e., large packet traces, their approach is not suitable for a fast estimation regarding the behavior of the observed RA scheme within the WLAN cell.

7.3 DARA Principle

The goal of our data rate estimation scheme is to identify the rate selection behavior of STAs with stationary traffic and constant packet sizes. We aim to obtain the rate selection behavior for such STAs after observing them for a short period of communication denoted as Δt_{trial} which is just in the order of some beacon intervals. Note that this estimation does not assume any preliminary knowledge about the actual rate adaptation scheme applied on each STA.

Basically, DARA consists of two different major blocks, the *acquisition of 802.11 parameters* and the *machine learning (ML) part* as shown in Fig. 7.1. The first, the acquisition of parameters[2], includes measurements of selected parameters on STA and AP side, their signaling as well as their preprocessing. On the basis of these 802.11 parameters,

[2]Throughout this chapter, we follow the notion of a parameter as an 802.11 system measure similar to Dujovne et al. [182].

the ML part conducts a *training* of a selected ML model. Once this training is completed, we use the ML model for *estimations*, again based on measured 802.11 parameters.

In the following, we first describe the outline of the approach. The major components of the selected ML model are described in detail in the next section, separately. The signaling of the 802.11 parameters is given in the final remarks in Sec. 7.8.

7.3.1 Outline of the Approach for an Individual RA Scheme

With DARA we aim to obtain the data rate selection of a given STA. Therefore, DARA uses the following abstraction of an RA scheme

$$f : \boldsymbol{x} \mapsto \boldsymbol{y}, \tag{7.1}$$

where $\boldsymbol{y}$ is the *output vector*[3] that gives us the probability mass function (pmf) of data rates resulting in successful uplink transmissions of a given STA. We consider $\boldsymbol{y}$ for a large time span $\Delta t_{\text{long-term}}$ being in the order of several tens of seconds. Note that $\boldsymbol{y}$ has an element for each data rate regarding its probability to be used for successful uplink transmissions. This results in the vector $[y_1, \ldots, y_{\text{R}}]$, where R gives the number of data rates for a given PHY (e.g., 802.11g ERP-OFDM offers eight data rates ranging from 6 to 54 Mbps). Accordingly, also f is a vector-valued function given by $f(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_{\text{R}}(\boldsymbol{x})]$. We refer to f as the *mapping function* in the following.

Regarding the inputs of DARA, the *input vector* $\boldsymbol{x} = (\boldsymbol{x_a}, \boldsymbol{x_b})$ consists of the two components $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$, both calculated over a small Δt_{trial} instead, which is just in the order of some beacon intervals. For $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$, we select technological parameters from the 802.11 PHY and MAC, discussed in the following and listed in Table 7.1.

Selected 802.11 Parameters

The first, $\boldsymbol{x_a}$, contains the mean values as well as the standard deviations of three 802.11 measures regarding a specific STA. The first is the RSSI of STA's uplink data frames successfully received by the AP. The second is the number of transmission attempts for each uplink data frame. The third is the "channel utilization" [6], which is the occupied airtime on the channel, seen at AP. Further, $\boldsymbol{x_b}$ is a small sample of $\boldsymbol{y}$ over the short Δt_{trial} aiming at potential internal effects of RA schemes which are not easily obtainable from off-the-shelf WLAN equipment. What we denoted as *processing* in Fig. 7.1 is essentially the computation of mean values, standard deviation, and pmf for the corresponding 802.11 parameter.

Note that for our observations, we stick to simple 802.11 parameters being available and easily accessible on STA or AP side [182]: *RSSI, successful data rates, transmission attempts, and channel utilization.* Acharya et al. [191] have shown that a channel utilization measure is suitable to reflect the load situation and to relate it with resulting packet losses in a WLAN cell, thus allowing us to identify the operational point of the network from the load perspective.

[3] We use the bold notation to emphasize vectors.

Table 7.1: Components of DARA's input vector

802.11 parameters	form of representation			measured at	
	mean	standard deviation	pmf	AP	STA
data rates of STA transmissions			✓	✓	
RSSI of STA's uplink transmissions	✓	✓		✓	
channel utilization	✓	✓		✓	
transmission attempts per data frame	✓	✓			✓

Training and Estimation Processes

The issue regarding the mapping function $f(\boldsymbol{x})$ given in Eq. 7.1 is that it is completely unknown. Thus we first need to derive a function $\tilde{f}(\boldsymbol{x})$ by a *training* process. This is realized by observing the instances $\boldsymbol{x}$ and $\boldsymbol{y}$ over the time span Δt_{train} consisting of multiple, subsequent Δt_{trial}. While we compute one set of input parameters $\boldsymbol{x}$ for each short Δt_{trial}, $\boldsymbol{y}$ includes the rate selection probabilities for successful transmissions over the whole observation period Δt_{train}. We choose $\Delta t_{\text{train}} \gg \Delta t_{\text{trial}}$ and $\Delta t_{\text{train}} < \Delta t_{\text{long-term}}$, whereby we select Δt_{train} such that it already gives us approximately a 'long-term' distribution of the successfully applied data rates.

We learn the mappings of different observed $\boldsymbol{x}$ and $\boldsymbol{y}$ pairs by means of an ML approach of *Gaussian processes* described in detail in Sec. 7.4. Once the model has learnt the mappings, we use it for *estimations*, i.e., we calculate the estimate $\boldsymbol{y}_*$ for a new input vector $\boldsymbol{x}_*$ obtained over a single period Δt_{trial}.

7.3.2 Handling Multiple RA schemes

Note that different RA schemes may lead to different mapping functions. We build classes of WLAN devices that are highly probable to apply a similar adaptation scheme. This allows us to have an abstraction for each class of WLAN NICs. The classes of devices are identified by their triple ⟨*vendor ID, NIC model, firmware version*⟩ as most likely the same RA scheme is applied on similar NICs.

Once we obtain all data over Δt_{trial}, we check whether DARA has been trained already with data from a NIC with the triple ⟨*vendor ID, NIC model, firmware version*⟩. If so, it loads the triple-specific parameterization for the ML model and the estimation for $\boldsymbol{y}_*$ on the basis of the input parameters $\boldsymbol{x}_*$ starts.

When a device with an unknown triple, i.e., NIC class, appears in the WLAN hotspot, no predictions are possible as the model is yet untrained. In this case, the AP collects measurements of this STA over long time spans Δt_{train}. Afterwards, we conduct the training for the concerned NIC.

For the sake of completeness, we also discuss at this stage a brief idea for an extension which has however *not* been modeled or evaluated throughout this work. If a device

belongs to a known class of NICs but the actual estimation for $\boldsymbol{x}_*$ delivers a high variance, i.e., the ML model itself has been pretty unsure about the quality of its output, we propose to take device's behavior for the re-training of the existing model. For this case, we suggest that the AP collects further measurements for such a STA and uses this data for retraining of the concerned NICs.

7.4 Machine Learning Model Used by DARA

For a training and an estimation for DARA, we make use of a ML model from the area of *Gaussian processes (GPs)* relying on *Bayesian methods* [192, 193]. While this is a research area by itself, we aim to guide the reader through our design choices in this section, while keeping mathematical sophistication at a minimum level. As a starting point, we consider the simple linear regression and its extension for non-linear relationships between model inputs and *one-dimensional* outputs y. On this basis, we discuss related issues being typical for ML applications. Then, we present our design requirements for DARA, leading to the choice of a model from the GP area. We introduce the notion of GPs and finally present all the details about the DARA model.

7.4.1 Linear Regression Revisited

Let us start the discussion about the selected model by considering the linear regression as illustrated by Bishop [192, Ch. 1 and Ch. 3]. For the sake of simplicity in the following example, let us assume that we have a one dimensional input, i.e., a single input parameter x. Then, the simple *linear regression model* is given by[4]

$$y(x, \boldsymbol{\omega}) = \omega_0 + \omega_1 x, \tag{7.2}$$

whereby $\boldsymbol{\omega}$ represents the vector with the weights[5]. Note that the function y has a linear relationship to the input x, so that this regression approach is only applicable if this linear condition is fulfilled.

To allow models also for cases where input and output have a non-linear relationship, the linear regression model from Eq. 7.2 may be transformed to a version with *basis functions* $\phi_j(x)$ [192]

$$y(x, \boldsymbol{\omega}) = \omega_0 + \sum_{j=1}^{M-1} \omega_j \phi_j(x), \tag{7.3}$$

whereby M gives the number of weights. Note that with $\phi(x) = x$ and $M = 2$, this model becomes again the simple linear regression from Eq. 7.2. In contrast, dependent

[4] In the examples we use $y(x)$ instead of $f(x)$ to differentiate between the application of DARA and more general regression / ML examples.

[5] In the ML literature, e.g., in [192], these weights are denoted as *parameters*, while the inputs are referred to as *variables*. To be consistent with our terminology from an 802.11 view, we contrary refer to the input as parameters and denote $\boldsymbol{\omega}$ as the weights.

on the specific relationship between x and y, one may select suitable basis functions of higher order that are not linear anymore, e.g., exponential or power functions.

Applying a model following Eq. 7.3 requires to select proper basis functions, to handle the order M of the basis functions, and to adapt the weights $\boldsymbol{\omega}$ to the training data. To deal with the latter issue, the simple method of *least squared errors* between the model output and the correct value of y can be used to tune the weights $\boldsymbol{\omega}$ [192].

Handling the order of the basis functions is not that easy, as an improper selection may result in the *overfitting problem*. Overfitting evolves, if one further and further increases the order of the basis functions, e.g., the degree of a polynomial model with $\phi(x)$ consisting of power functions. At a certain instance, then the model will match all the given points of available training data very well, but produces large prediction errors for other x. As a matter of fact, Bishop reports as a rule of thumb that the number of training points should be 5 to 10 times higher than the number of weights $\boldsymbol{\omega}$. This however leads to the unfavorable situation that the model complexity just directly depends on the amount of collected training data and not on the actual considered learning problem [192].

7.4.2 Requirements for the DARA Model

By means of the basic considerations from the last section, we now formulate our requirements for the selection of the DARA model pointing out the differences regarding the simple regression. We shall note that all the requirements discussed below may be fulfilled with a ML model using Gaussian processes.

Unlike the previous regression case, we do not want to formulate an explicit function $f(\boldsymbol{x})$ of a rate adaptation scheme. In other words, we are not interested in the exact description of the functional behavior between $\boldsymbol{x}$ and $\boldsymbol{y}$, as we aim to avoid a selection of basis functions or the determination of the number of weights. In line with this, we do not want to take care about the overfitting problem. Instead, we aim at a model that is capable to handle its complexity of the weights automatically.

As mentioned above, DARA takes 14 different WLAN parameters as input. Regression with basis functions may impose significant issues if the dimension of inputs becomes large [192, pp. 33–38]. As such, we would be required to classify parameters into significant and non-significant subsets, essentially reducing the dimension of the inputs. Instead, we aim for a model that is capable to identify significant inputs itself.

Lastly, besides $\boldsymbol{y}$ we prefer an additional output of the model regarding its 'certainty' for an estimation at a given point. With this, we aim to decide for parameter combinations, where our model may require a retraining because its prediction is not certain enough and too 'noisy'. It is especially this feature that is unique for Gaussian processes. Sec. 7.4.4 below discusses the selected GP model in detail.

7.4.3 Definition of a Gaussian Process

Before we finally come to the specifics of the selected ML model, let us first describe what a *Gaussian process* is and how it is related to our learning problems, again for a

one-dimensional output y. Let us start with a simple basis: The well-known *Gaussian distribution* for one random variable x is given by [192, 193]

$$x \sim \mathcal{N}(\mu, \sigma^2), \tag{7.4}$$

with mean μ and variance σ^2. This distribution can be extended to a Gaussian distribution over *multiple* random variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$. The resulting *multi-variate Gaussian distribution* is given as

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), \tag{7.5}$$

whereby Σ is the co-variance matrix between $x_1, x_2, \ldots, x_N$. Note that Gaussian distributions specify random variables (for the one-dimensional case) or vectors (for the multi-variate case).

Taking a step further, "a Gaussian process is defined as a probability distribution over functions $g(\boldsymbol{x})$ such that the set of values $g(\boldsymbol{x})$ at an arbitrary set of points $x_1, ..., x_N$ jointly have a Gaussian distribution" [192, p. 305]. Then, functions $g(\boldsymbol{x})$ are distributed as [193]

$$g(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x_i}), k(\boldsymbol{x_i}, \boldsymbol{x_j})), \tag{7.6}$$

where $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ denote arbitrary input vectors, while $m(\boldsymbol{x_i})$ and $k(\boldsymbol{x_i}, \boldsymbol{x_j})$ are the *mean* and the *covariance function*. The latter part is of great importance for our learning problem, as it completely gives us the properties of the distribution over the functions $g(\boldsymbol{x})$. Now, the trick with Gaussian processes is that one just operates on preselected covariance functions of the input vectors, thus determining the statistical properties of the distribution, without actually deriving $g(\boldsymbol{x})$ explicitly.

7.4.4 The Selected Gaussian Process Model

For DARA, we assume that the rate adaptation algorithm is unknown and consequently the form of our mapping function $f(\boldsymbol{x})$ is not available. And even in those cases where the rate adaptation algorithm is open source, that does not help much: the relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the result of complex dynamics and could only be calculated by simulating it stepwise. Therefore we emphasize again that we do not try to model this function explicitly, but use a non-parametric approach to learn the complete mapping from $\boldsymbol{x}$ to $\boldsymbol{y}$. For that purpose, we assume that a function $g(\boldsymbol{x})$, covering all data rates, is a sample from a *GP*. As mentioned above, such a GP is a distribution over functions leading to the expectation values [192]

$$E[g(\boldsymbol{x_i})] = 0 \quad \text{and} \quad E[g(\boldsymbol{x_i})g(\boldsymbol{x_j})] = k(\boldsymbol{x_i}, \boldsymbol{x_j}), \tag{7.7}$$

where $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ denote arbitrary input vectors. As explained by Bishop [192], any finite set of function values $\{g(\boldsymbol{x_1}), g(\boldsymbol{x_2}), \ldots\}$ comes from a multinomial Gaussian distribution fully defined by the first and second moments given in Eq. 7.7.

We extend this approach to multidimensional outputs, like our data rate distribution $\boldsymbol{y}$, by adding an input value which selects the desired output component. This

input value is an index r ranging in between $1 \dots \mathrm{R}$, whereby R gives the number of data rates for an 802.11 PHY. For example, for the eight 802.11g ERP-OFDM data rates of 6 to 54 Mbps, we compute for each data rate $r = 1 \dots 8$ a corresponding y_r, such that $\boldsymbol{y}$ finally consists of $[y_1, \dots, y_8]$.

The kernel $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \mathrm{Cov}(g(\boldsymbol{x_i}), g(\boldsymbol{x_j}))$ describes variance and autocorrelation of the function $g(\boldsymbol{x})$, which are easier to specify than a parametric form. For our model we choose the squared exponential kernel with *automatic relevance determination (ARD)*

$$k_1(\boldsymbol{x_i}, \boldsymbol{x_j}) = \theta_0 \exp\left[-\frac{1}{2}\sum_{m=1}^{D} \eta_m (x_{i,m} - x_{j,m})^2\right].$$

Here $\theta_0, \eta_1, \dots, \eta_D$ are *hyperparameters*, which can be adapted to the training data set $B = \{\boldsymbol{x_i}, y_i | i = 1...N\}$, with D being the dimension of the input vector $\boldsymbol{x_i}$ and N being the number of observations. High values of the weight η_m indicate that the m-th element of x is relevant, while its influence on y is small for low values [193]. We also take into account that the model might not fit perfectly and assume Gaussian observation noise with variance θ_2:

$$k_2(\boldsymbol{x_i}, \boldsymbol{x_j}) = \theta_2 \, \delta_{i,j}.$$

Linear effects are modeled explicitly by a linear kernel

$$k_3(\boldsymbol{x_i}, \boldsymbol{x_j}) = \theta_3 \sum_{m=1}^{D} x_{i,m} \, x_{j,m}$$

with hyperparameter θ_3. Finally, the full kernel used in the ML model is given by the sum

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{n=1}^{3} k_n(\boldsymbol{x_i}, \boldsymbol{x_j}).$$

As we do not know the correct values of the hyperparameters $\theta = (\theta_0, \eta_1, \dots, \eta_D, \theta_2, \theta_3)$ before, these are adapted to B applying standard Bayesian methods [192] leading to a minimization of the negative marginal likelihood

$$-\log p(\boldsymbol{y}|\boldsymbol{x}, \theta) = \frac{1}{2}\boldsymbol{y}^\top K_B^{-1} \boldsymbol{y} + \frac{1}{2}\log|K_B| + \frac{n}{2}\log 2\pi, \tag{7.8}$$

where K_B is the covariance matrix between the y contained in B given by $(K_B)_{i,j} = k(\boldsymbol{x_i}, \boldsymbol{x_j})$ [193]. Note that the Bayesian viewpoint already includes a tradeoff between the data fit of a model and its complexity, thus handling the overfitting problem automatically [193].

Finally, after minimizing the negative log likelihood, it requires only linear algebra to perform GP regression [193]. Thereby, the estimate y_{r_*} for a specific data rate r is given by

$$y_{r_*} = E[g(\boldsymbol{x_*})] = \boldsymbol{k_*}^\top K_B^{-1} \boldsymbol{y},$$

where $\boldsymbol{x}_*$ denotes the test input and $\boldsymbol{k}_*$ is the vector of covariances between $\boldsymbol{x}_*$ and B. We also get the variance

$$\sigma_{r_*}^2 = \mathrm{Var}(g(\boldsymbol{x}_*)) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\top K_B^{-1} \boldsymbol{k}_*$$

of the function at the test point, which is an indicator for the uncertainty of the prediction. This provides an advantage compared to non-Bayesian methods, where confidence intervals are more difficult to determine.

7.5 Selected Settings for DARA's Evaluation

For a 'proof of concept' of DARA, we consider two real RA schemes—AARF/AMRR and Minstrel. For both, we are interested in the obtainable accuracy for estimations if we train DARA with data of random durations Δt_{train}. We evaluate the accuracy of estimates by considering the mean squared error between predictions and real values. For this work, we study estimates for VoIP traffic as its strong QoS requirements make it interesting for a support of timely decisions, e.g., for handovers or WLAN reconfigurations. Large-scale measurements showed a significant fraction of calls having a duration above 25 s [194]. To account also for larger call lengths, we select calls for the training such that Δt_{train} uniformly distributes between 20 and 40 s. For $\Delta t_{\mathrm{trial}}$, we use ten consecutive 100 ms beacon intervals observing the short-term RA behavior for up to 50 VoIP packets.

Methodologically, we first considered simulations by ns-2 with detailed SINR and fading models as described in Appendices B.3 and B.4. In the simulations, we did not consider other active WLAN hotspots on interfering channels. Second, we conducted a campaign of measurements in our WLAN testbed located in our office environment on campus, to also cover interference as a result of high hotspot densities in today's WLAN channels (again, since it is not covered by simulations).

This section gives a short overview over the selected RA schemes, elaborates on the simulation and testbed setup in which we obtain the data for training and estimation, and describes the conducted training process with its tool chain.

7.5.1 Selected Rate Adaptation Schemes

Both in simulations as well as measurements, we applied *AARF/AMRR* and *Minstrel*. From the RA schemes included in the Linux 802.11 WLAN driver MadWifi, Minstrel was shown to be the most sophisticated scheme [187, 190]. In addition, we consider also the popular family of AARF/AMRR schemes, which was shown before to struggle with congested environments [188].

The AARF scheme is an extension of the *auto rate fallback (ARF)* algorithm, which increases the data rate after ten acknowledged data transmissions and reduces its rate if two contiguous transmission attempts remain unsuccessful. As this static adaptation of ARF has been shown to be susceptible to fluctuating wireless channels, AARF dynamically tunes the number of contiguous successful and non-successful transmissions for up- and downgrading the data rates in addition to the ARF operation. AMRR is a specific

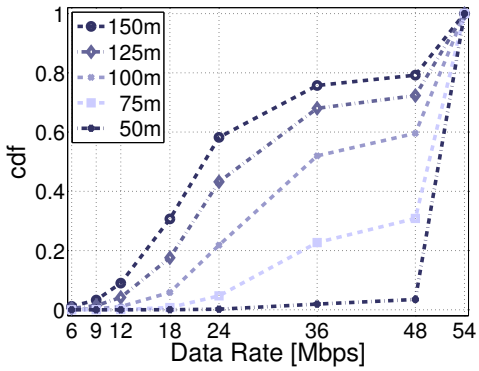


Figure 7.2: Distribution of data rates with 'perfect' rate selection

flavor of AARF enabling to deal with the specific timing behavior of the Madwifi driver. While in our simulations, we stick to the original definition of AARF [173], we apply AMRR in our measurements as it comes originally with the Madwifi driver [186].

The second scheme, Minstrel [195], conducts a certain fraction of transmissions for testing suitable, alternative data rates and observes all transmission attempts. More specifically, it records the number of successful and failed transmission attempts appearing at each data rate. From this, the algorithm derives a success probability smoothed by an exponentially weighted moving average (EWMA), and estimates on this basis the expected throughput for each data rate. Finally, it selects the rates with the highest expected throughputs. Periodically, Minstrel evaluates the success probability and the estimation of the expected throughputs each 100 ms.

7.5.2 Setup for the Simulation-based Training

In our simulations, we trained the ML model at an operational point of the network at which no further VoIP flows can be accommodated in the WLAN cell (QoS constraints and simulation settings are given below); this results in about 40 VoIP nodes transmitting concurrently in the hotspot. Thereby, the first randomly selected half of nodes applies AARF, while the rest operates Minstrel. Further, the AP uses Minstrel for the downlink transmissions in our setting.

Simulation Model

For our simulations, we used the same models as given in Sec. 6.6.4. In summary, we used an IEEE 802.11g AP that is 11e-capable by providing EDCA functionality. All WLAN devices applied 802.11g ERP-OFDM—from 6 up to 54 Mbps. Further, we applied a Log-Distance Path Loss, a Ricean Fading and a SINR model as given in Appendix B. Additionally, we extended the WLAN model by the two RA schemes AARF and Minstrel.

Population of Nodes, VoIP Traffic Model, and QoS Constraints

Within the joint coverage area of WLAN and WWAN, we assume to have 200 VoIP nodes. For the generation of their VoIP traffic, we used the same model, parameterization, and QoS constraints as detailed in Sec. 6.4.5, i.e., an exponential ON/OFF model generating packets according to the G.711 codec (160 Byte audio packets each 20 ms during ON periods). Again we consider a hard QoS limit for VoIP in terms of losses, consisting of lost and late packets. If five or more percent of the VoIP packets are lost over the last ten seconds, the quality is assumed to be poor.

Scenarios

Throughout the whole chapter, we investigate five scenarios in which we vary the size of the quadratic area with edge lengths of 50, 75, 100, 125, and 150 meters; the AP resides at a corner of this area; the STAs are random uniformly distributed. Larger edge lengths thereby lead to greater sizes of the areas such that the data rates applied for 802.11 transmissions are reduced. Fig. 7.2 shows the cumulative distribution function (CDF) of data rates for each edge length, if the rate selection is done on ideal 'perfect' knowledge about the rate to be applied.

Filling the Hotspot with VoIP Nodes

We apply the following basic algorithm with which we model a selection of VoIP users and the operation of an admission control for the WLAN hotspot. We evaluate periodically each three seconds whether the QoS values of the active STAs in WLAN have been violated. In case of harmed QoS limits, we select randomly one of the STAs with bad QoS and trigger a handover to WWAN by means of IEEE 802.21. In contrast, if no QoS violation has occurred, we accommodate an additional user in the WLAN cell by selecting him in a random-uniform fashion from all STAs being within WLAN coverage as well as being served by the WWAN.

Selection of Training Nodes

As we want to cover high- as well as low-rate STAs, we trained the model on the scenario where the 'worst rate' distributions appear, i.e., the large scenario with 150 m edge length. The motivation for this is a follows: by the simulations we aim to cover a broad range of different SNR regions. Thus, for the training of our model, we assigned nodes to bins with radii around the AP of multiples of 30 meters. For the scenario with an edge length 150 m, we selected nodes for a training from 5 bins, namely from the second to the sixth bin. We omitted nodes from the first bin as we targeted to focus on SNR regions, where rate changes are likely to appear as a result of SNR variations. Regarding the first bin, Fig. B.3 in Appendix B.4 shows that positions close to the AP on average have very strong SNRs far above the thresholds of the data rates. Thus, from the second to the sixth bin, we randomly selected one NIC of each type (either AARF or Minstrel)

from each bin as a training node. Fig. 7.3 gives the data rate pmfs obtained from the training data for the different node positions and the RA schemes.

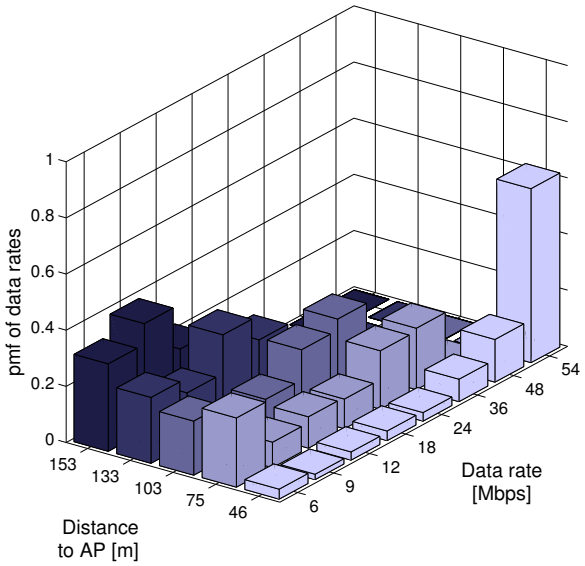
7.5.3 Setup for the Testbed-based Training

In addition to the simulations, we verified our approach in a testbed on campus, which also includes interference from other WLAN hotspots in overlapping channels. For the training with testbed data, we followed a two-stage methodology as pure VoIP traffic leads to a low-loaded WLAN with our available hardware. First, we trained our ML model with data observed in pure VoIP scenarios. Second, we considered measurements from scenarios with VoIP together with background traffic of TCP downloads by a STA, e.g., resulting from HTTP and FTP downloads or TCP video streaming solutions. In all measurements, all VoIP nodes together applied the same RA scheme, either AMRR or Minstrel. In contrast, the AP and the STA conducting TCP downloads used Minstrel.

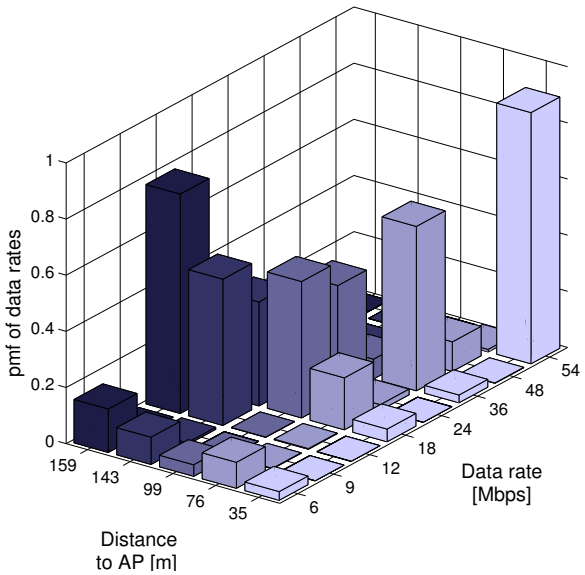
Testbed Environment

The measurements were conducted in three different rooms on the first floor (Fig. 7.4) of our building located on campus. The university itself operates a large campus-wide WiFi network to offer wireless access. As a result, our measurement environment had to deal with interferences from different floors as well as from neighboring buildings. We operated our testbed on WLAN channel 1; during the measurements we were seeing between 15 and 20 other WLAN hotspots within channels 1 to 5.

In the testbed, we applied four WLAN STAs and one AP, all of them running on laptops of the type IBM R50/R51 or Lenovo R500 (CPUs of 1.5 to 2.53 GHz and 768 MB to 2 GB RAM). They include CardBus WLAN NICs with Atheros chipsets (Netgear WAG511 v1/v2 and WPN511) all driven by the Madwifi version including, at this particular time, the most recent *hardware abstraction layer (HAL)* [196]. In our settings, this helped avoiding Madwifi's well-known HAL stability problems [197] over large time scales from a couple of hours up to several days. For all nodes, transmit power was fixed to 10 dBm, power save as well as RTS/CTS was turned off, and 802.11g ERP-OFDM with 6 to 54 Mbps was applied. Further we used 802.11e EDCA with the VoIP (AC_VO) QoS class and the Background (AC_BK) class for TCP (with Madwifi's standard settings for 802.11e parameters [6] of BK: AIFSN = 7, CW_{min} = 15, CW_{max} = 1023, and VO: $AIFSN_{AP}$ = 1, $AIFSN_{STA}$ = 2, CW_{min} = 3, CW_{max} = 7). All notebooks were operated with Debian 6.0 and the vanilla kernel 2.6.32.46 without experimental modules. We wired the WLAN devices by an 1 Gbps Ethernet backplane over which we transport measured data to an SNMP manager located on a separate server. Further, the devices were synchronized with the *Precision Time Protocol (PTP)* [198, 199] for a precise, internal time stamping of measurement data.



(a) AARF



(b) Minstrel

Figure 7.3: Training pmfs of data rates from simulations

Collection of Data

On the sender side, we built a wrapper around the RA module tracing all in- and outputs, i.e., the rate selection, the number of retransmissions, information about the success of the transmission, the sequence number, and the timestamp for each data packet. On the receiver side, we log all events in the Madwifi driver that arrive at the routine handling successfully received data packets. There, we trace the applied data rate, the packet size, the RSSI value, the sequence number, and the according timestamp. At the AP, we trace the occupied airtime of all WLAN traffic in up- and downlink direction per beacon interval of 100 ms by periodically reading out the related registers as described by Acharya et al. [191].

Placement of WLAN Nodes

Fig. 7.4 shows a snapshot of the RSSI values from the hotspot in our office environment obtained with the software "Netspot" [200]. Within the labeled rooms *R1* to *R5*, we measured the RSSI-values from our hotspot at least at 20 different positions per room. Although this is an exemplary snapshot, it helped us understanding the propagation of the radio signals in our testbed, where we placed our AP at the upper right corner of room 1. Repeating these snap-shots on different days showed that rooms 3 and 4 offer the most interesting positions as they show strong fluctuations in RSSI within small areas. We aim to consider such good and bad RSSI conditions by our training to include best and worst cases. Further, in contrast to the training of the simulations, where we covered a multitude of RSSI regions, for the measurements we aim to keep the number of devices small. Thus, we just considered one additional device with RSSI conditions ranging between the 'good' and 'bad' positions. As a result, we conducted the following placements of WLAN STAs: our 'bad' STA is put in room 4 in the low RSSI region, the 'medium' STA resides in room 3, while the 'good' STA stays in room 1 together with the AP. Further, we placed our TCP background-load (TCP-BG) STA in room 1 close to the good STA.

Each WLAN node in room 1, 3, and 4 was placed at two different positions (set 1 and 2) with a distance of about 3.5 m to each other. To separate the training of the ML model from the actual estimation, we used the first set of positions for the training and data from the second set for estimations, only.

Traffic Model

We emulated VoIP traffic by generating packets with a 160 Byte audio payload each 20 ms by means of the traffic generator tool "Iperf" [201]. The same QoS limits as in Sec. 7.5.2 apply for our VoIP traffic. Further, we generated TCP-BG traffic with TCP downloads. During all measurements, our TCP-BG STA ran a download at different goodputs from a server within our wired infrastructure. For this we applied an FTP-Server on the basis of "Pure-FTPd" [202] and initiated the downloads by "Wget" [203] on STA side. The duration of a complete download always exceeded the measurement duration in every setting. To study different operational points of the WLAN hotspot,

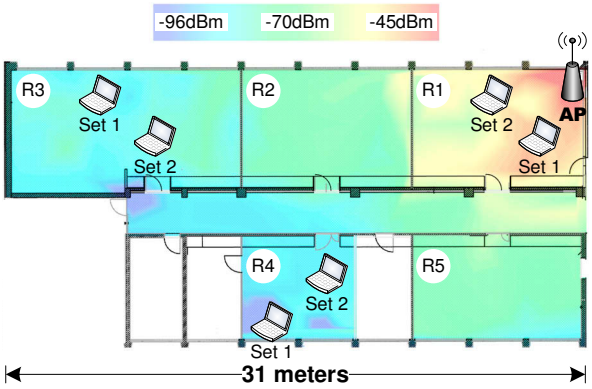
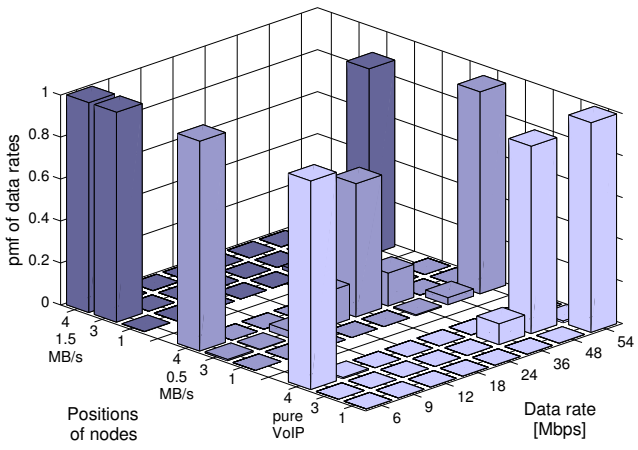


Figure 7.4: Testbed environment

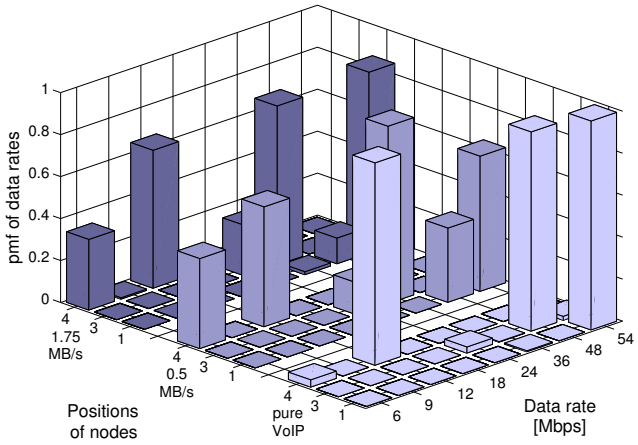
we varied the goodput of the TCP-BG STA by throttling the download on server side. We trained AMRR and Minstrel both with pure VoIP and different TCP-BGs. For AMRR, TCP-BG goodput was set to 0.5 and 1.5 MB/s, while Minstrel training was conducted with 0.5 and 1.75 MB/s. The high goodput values led to operational points of the hotspot such that medium and bad STA started to perceive QoS degradations in some Δt_{train}. Fig. 7.5 shows the data rates pmfs that were used for the training of the model. In the figure, each node number refers to the room in which it is placed on its 'set 1' position during the training.

7.5.4 Training the Machine Learning Model

As described in Sec. 7.4, the hyperparameters of the ML model needed to be adapted, i.e., optimized, for the collected training data $\boldsymbol{x}$ and $\boldsymbol{y}$. We fed $\boldsymbol{x}$ and $\boldsymbol{y}$ into the "gpml-framework" [204] which calculates the negative log likelihood from Eq. 7.8 for the selected kernel and its hyperparameters. We obtained optimal hyperparameters separately for simulation and measurement data by minimizing the negative log likelihood with Matlab's optimizer "fminunc" [205] which applies a "trust-region method" [205, 206] on the basis of the "interior-reflective Newton method" [207]. To ensure that we found a (local) minimum in the negative log likelihood, we consider the gradient (being sufficiently close to zero) and the eigenvalues of the Hessian matrix (being all greater than zero).



(a) AMRR with pure VoIP and TCP-BG of 0.5 and 1.5 MB/s



(b) Minstrel with pure VoIP and TCP-BG of 0.5 and 1.75 MB/s

Figure 7.5: Training pmfs of data rates from measurements

7.6 Evaluation of DARA's Accuracy

To judge the accuracy of DARA's estimators obtained in Sec. 7.5 for AARF/AMRR and Minstrel, we conducted estimations for new input parameters $\boldsymbol{x_*}$. For this, we shifted our WLAN devices to different positions (in our measurements) or observed a disjoint set of WLAN nodes that had not been used for obtaining training data (in simulations). Further, we ensured that our data sets for estimations do not overlap in time with data used for training. By this, we obtain $\boldsymbol{x_*}$ for which DARA was not trained.

7.6.1 Accuracy Metric

At each node selected for an estimation, we conducted 32 estimations $\boldsymbol{y_*}$ and compare the estimated data rate distributions with the real values $\boldsymbol{y}$. As a metric for the accuracy, we consider the mean squared error (MSE) for each individual data rate r ($\mathrm{MSE}_{\mathrm{rate_r}}$) [193]

$$MSE_{\mathrm{rate_r}}(y_{\mathrm{r}_*}, y_{\mathrm{r}}) = \frac{1}{N} \sum_{i=1}^{N=32} (y_{\mathrm{r}_*} - y_{\mathrm{r}})^2.$$

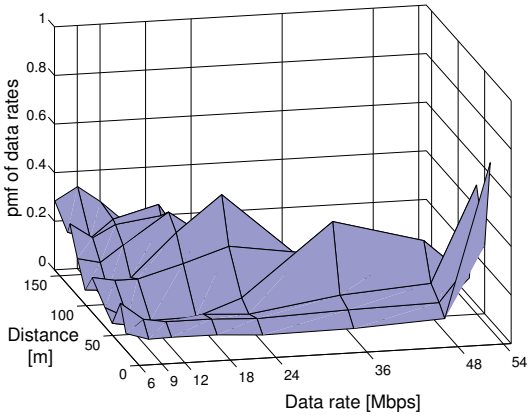
Note that the MSE specifically penalizes large differences between y_{r_*} and y_{r}. Further, to enable also a visual comparison of y_{r_*} and y_{r} directly, we calculated at each node position and for each individual data rate the mean value over all repetitions. Their confidence intervals with a 95 percent level stayed within $+/-0.06$ around the mean, if not stated otherwise. Thus we do not plot them for better visibility.

7.6.2 Simulation Results

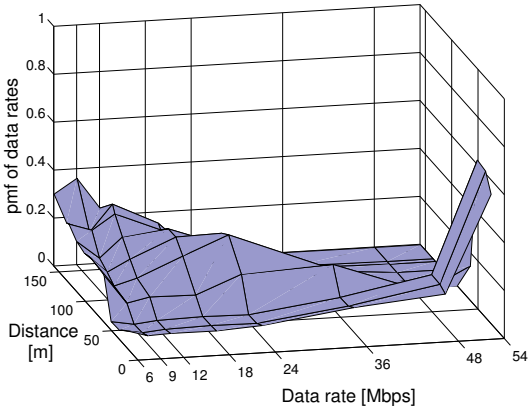
In order to enable an evaluation of STAs operating in different MCS regions, we considered the simulation scenario with an edge length of 150 m. For each of the selected rate adaptation schemes, we selected 10 test nodes with radii of multiples of 15 meters around the AP to cover all MCS regions.

The results are shown in Fig. 7.6 and 7.7[6] for AARF and Minstrel, respectively. The values are plotted over the data rates and for the different positions of the selected test nodes. Both plots show the mean measured and the mean estimated values as well as the MSE over the 32 repetitions. These results show that the AARF estimates perform on average well, but impose certain errors at high rates for near nodes and smaller errors at medium rates for middle positioned nodes. Considering Minstrel estimates, the similarity between the curves of the measured values (Fig. 7.7a) and the estimated values (Fig. 7.7b) is also dominant. We note that for the mean estimated values in Fig. 7.7b for the two closest STAs to the AP, we obtain slightly larger confidence intervals of $+/-0.08$ and 0.09 for the high data rates of 54 Mbps. The MSEs in Fig. 7.6c show that additional prediction errors appear for the furthest nodes transmitting at low rates.

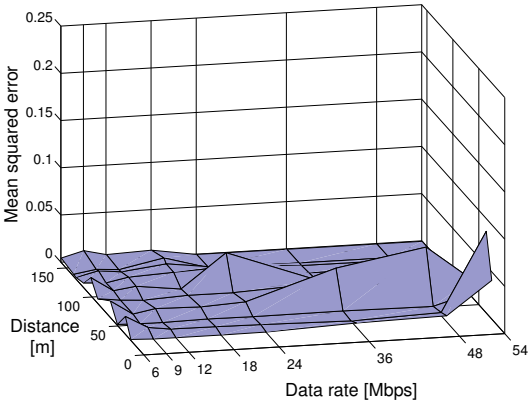
[6] We are aware that measurements, estimates and MSEs of different data rates do not obey a continuous function. Nevertheless, to allow an intuitive graphical presentation of such a high number of data sets, we interpolate the data points.



(a) Measured values

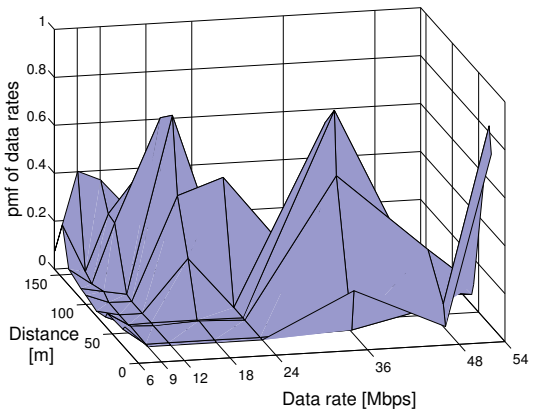


(b) Estimated values

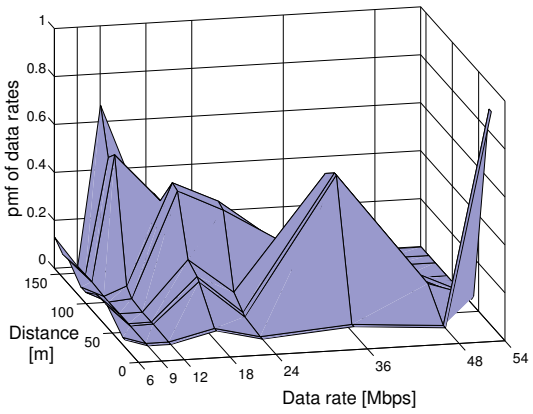


(c) Mean squared error

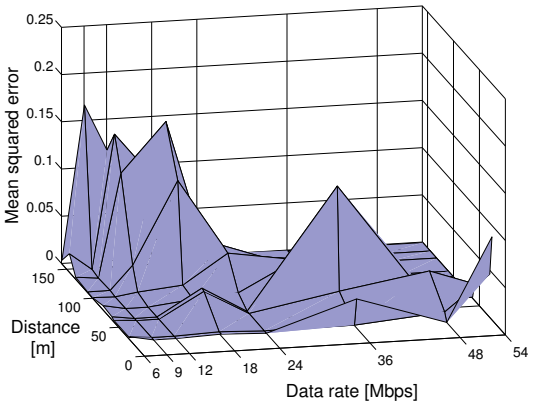
Figure 7.6: AARF: simulation results



(a) Measured values



(b) Estimated values



(c) Mean squared error

Figure 7.7: Minstrel: simulation results

Note that these errors are conservative in the sense that the predictions include lower rates than contained in the actual, measured values.

7.6.3 Results from the WLAN Testbed

In our testbed, we conduct all measurements for the estimation from the second positions (set 2 in Fig. 7.4) and with TCP-BG loads of 0.25, 0.75, 1.25, and 2.0 MB/s (denoted as low TCP1 and 2, medium TCP, and high TCP). The data for estimations have been obtained a couple of days after the training measurements (Sec. 7.5). Typical for highly utilized WLAN channels, our environment also shows strong, time-varying interferences from other WLAN APs.

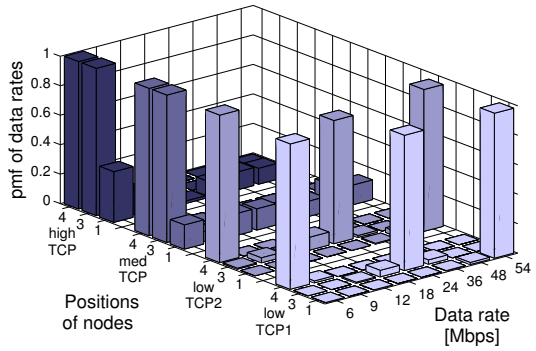
Results for the estimation in our testbed are shown in Fig. 7.8 and 7.9 for AMRR and Minstrel. In the figures, each node number refers to the room in which it is placed on its 'set 2' position. Note that for AMRR, at certain positions almost all transmissions appear either at 6 or 54 Mbps data rates, where we obtain very small confidence interval (0.01 and below). In contrast, with an increasing background load, the rate distribution starts to show increasing variations, specifically for node 1 and 3. Similar to the simulation results, considering MSEs, the estimations for AMRR show small errors for the medium positioned node 3. Some errors appear for node 1 at medium and high TCP-BG scenarios. This directly relates with high variations in the measured values of the corresponding STAs. Confidence intervals for node 3 at low TCP2 background load is $+/-0.09$ for 24 Mbps. For node 1 at medium and high TCP background loads, confidence intervals range from $+/-0.04$ to 0.014.

In contrast, Minstrel estimations in Fig. 7.9c show more and stronger MSEs than AMRR estimates. Smaller MSEs for low RSSI regions appeared in simulations already. However, in the testbed, large MSEs were observed for data rates above 18 Mbps, appearing pairwise mostly at 24 and 36 Mbps especially for the medium node 3. Actually, these pairwise errors result from a 'wrong' guess of the ML model regarding the exact data rate: Instead of correctly estimating a large fraction of transmissions at a certain rate, the model predicts this at the neighbor data rate—leading to large MSEs for both.

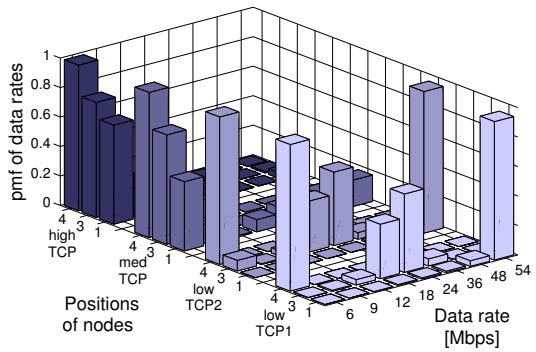
7.6.4 Conclusions from the Estimation Results

From the simulation and testbed results, we conclude that obtaining the exact pmf of the data rates is error prone for the sophisticated RA scheme Minstrel. Especially, in the testbed setup dealing with strong interference from other hotspots, the selected ML model is not completely capable of covering Minstrel's internal, statistical behavior for the high rates of 24 Mbps and above and thus mismatches a certain fraction of data rates with the direct neighbor rate. However, note that DARA does not have any knowledge about the internal behavior of an RA scheme and is only estimating on the basis of simple accessible PHY/MAC parameters described in Sec. 7.3.1.

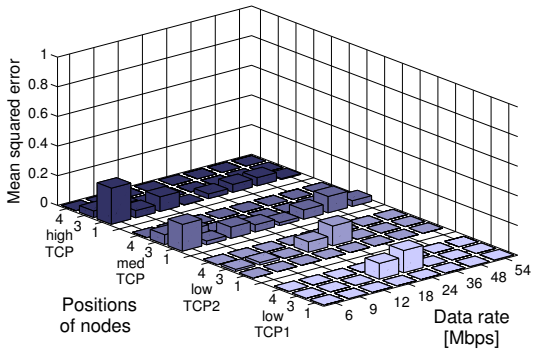
Thus, the results reveal that—although the exact pmf cannot obtained correctly in all cases—one is still able to determine a 'trend' for a considered STA, i.e., whether it



(a) Measured values

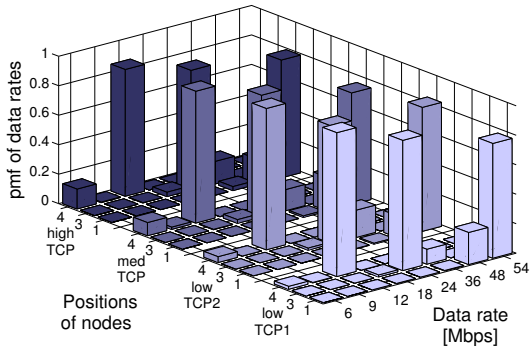


(b) Estimated values

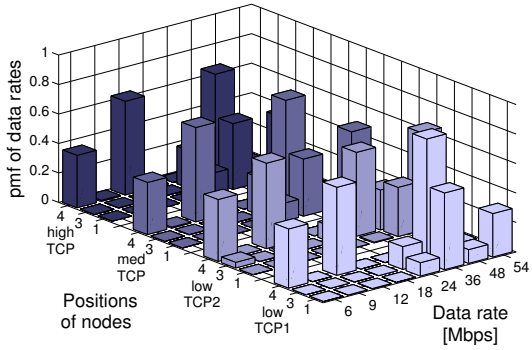


(c) Mean squared error

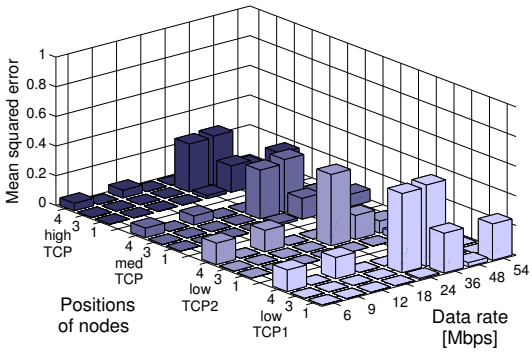
Figure 7.8: AMRR: measurement results with TCP-BG loads of 0.25, 0.75, 1.25, and 2.0 MB/s, denoted as low TCP1 and 2, medium TCP, and high TCP



(a) Measured values



(b) Estimated values



(c) Mean squared error

Figure 7.9: Minstrel: measurement results with TCP-BG loads of 0.25, 0.75, 1.25, and 2.0 MB/s, denoted as low TCP1 and 2, medium TCP, and high TCP

operates in the low-rate regions or in the higher-rate areas—being just dependent on the selected RA scheme. Such a trend may help in today's network in various applications.

7.7 Applying DARA for Technology Selection

In this section, we target the second and the third research question introduced in the beginning of this chapter. Aiming to avoid devices in our WLAN hotspot which transmit only in a low MCS region, we assume that a device subject for an offloading first associates with a target WLAN cell being determined by classical, rather heuristics-based network selection schemes. Then however, we observe the first transmissions of this device in WLAN afterwards. These transmissions could be, for example, related to the signaling of a handover or may even include the first data transmissions on the WLAN link. In this chapter, we focus on the latter aspect. On the basis of these observations, we make an estimation whether the newly associated device will be suitable in the WLAN cell on a long-term scale regarding its rate selection behavior. We define suitability as a fraction of low rates applied for transmissions. The exact parameterization is given further below.

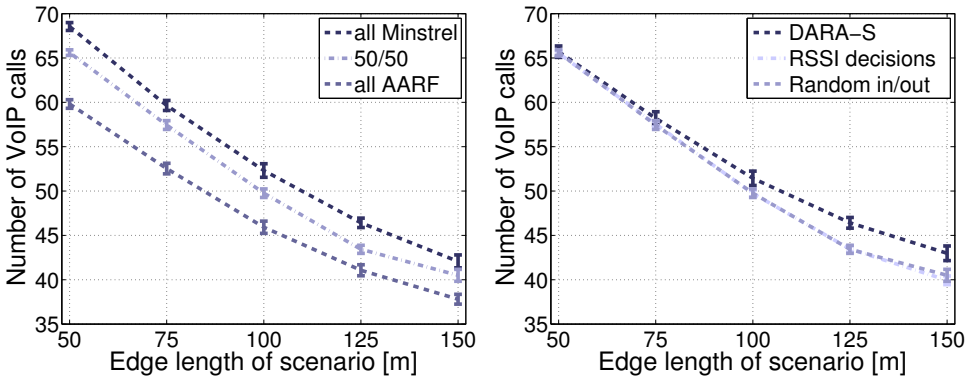
For our performance evaluation we used simulation settings already described back in Section 7.5.2. Now, we study improvements regarding the maximum VoIP capacity of a WLAN hotspot if one bases the selection of VoIP streams on the rate adaptation behavior of a STA. For this, we compare DARA with the common RSSI-based decisions.

7.7.1 Schemes for the Selection of WLAN Flows

We extend the basic algorithm given in Sec. 7.5.2 with which we model a selection of VoIP users and the operation of an admission control for the WLAN hotspot.

Again, the basic algorithm evaluates periodically whether the QoS values of the active STAs in WLAN have been violated. In case of harmed QoS limits, either for the new or other STAs in WLAN, we select randomly one of the STAs with bad QoS and trigger a handover to WWAN by means of IEEE 802.21. In contrast, if no QoS violation has occurred, we accommodate an additional user in the WLAN cell by selecting him in a random-uniform fashion from all STAs being within WLAN coverage as well as being connected to WWAN. We denote this basic scheme as *random in/out* in the following.

Now, the extension for this section covers a different handling of new users being accommodated by the WLAN cell on top of 'random in/out'. It works as follows. If not done before, the new user associates with the WLAN cell. Upon the start of his uplink transmissions, we observe his behavior over Δt_{trial} and determine whether to keep him in the WLAN afterwards. If we do not select him for WLAN access, we trigger a handover to WWAN, again by means of with IEEE 802.21. We study two different schemes to determine whether to hand over or to keep a user in WLAN—RSSI-based decisions as well as selections on the basis of DARA. We denote the latter as DARA-based selection (DARA-S) in the following.



(a) Mixes of RA schemes with 'random in/out' (b) DARA-S, RSSI, and 'random in/out'

Figure 7.10: VoIP capacity in different scenarios

RSSI-based Decisions

As discussed in Sec. 3.4.3, it is a common approach to base the network selection on the RSSI of a considered STA. For this, we collect the RSSI-measurements of STA's uplink data transmissions at AP over Δt_{trial}. Afterwards, we average the measured values and compare it with a threshold. This threshold has been selected such that STAs should be able to transmit the high rate part of their frames (on average) at least with 12 Mbps.

DARA-S

Here, we identify low rate STAs by means of DARA's estimation applying trained ML model from Sec. 7.5.2. To decide for the permanent selection of an arriving STA, we consider the probabilities of the low rates from the estimated histogram. Similar to the RSSI-based decisions, we target to avoid 6 and 9 Mbps transmissions. We trigger a handover, if the sum of probabilities for the *set of low rates*, i.e, 6 and 9 Mbps, is above 20 percent. We selected this threshold to allow RA schemes a 'testing' of low data rates for their internal statistics.

7.7.2 Performance Evaluation

In order to identify the gains of DARA-S, we finally conducted a set of simulations in which we studied the maximum number of VoIP clients in the WLAN cell.

There, we followed a two-stage methodology. First, we studied the impact of the different RA schemes. For this, we analyzed the VoIP capacity if no sophisticated selection scheme for arriving VoIP devices was applied, i.e., STAs are selected in a random fashion (by 'random in/out'). There, we studied three different mixes of rate adaptation algorithms: in the first two mixes, the AP and all STAs together either applied AARF or Minstrel (denoted as *all AARF* and *all Minstrel*). In the third one, the AP used

Minstrel while the (randomly selected) first half of STAs applied AARF and the rest Minstrel (denoted as *50/50 mix*). Then, in the second stage, we conducted our selection schemes on the 50/50 mix and compared it with the 'random in/out' selections. Lastly, we studied the accuracy of DARA-S's classifications.

For each STA mix, each decision scheme and each scenario size, we conducted independent replications [153], whereby the repetitions were ranging from 16 to 38 (all Minstrel: 19, all AARF: 20, 50/50 mix with pure 'random in/out': 26, DARA-S: 16, and RSSI decisions: 38 simulation runs). On this basis, we calculated confidence intervals with a 95 percent confidence level.

Mixes of RA Schemes

Fig. 7.10a shows the results for the mixes of STAs with different RA schemes. 'All Minstrel' outperforms all others, while the 'all AARF' case leads to the smallest number of VoIPs in all scenarios. The '50/50 mix' already leads to a significant improvement over 'all AARF'. These results show how different RA schemes influence the VoIP capacity of a WLAN cell. The distance between 'all Minstrel' and the '50/50mix' gives an indication about the potential gains for selection schemes taking the issue of the rate adaptation of STAs into account.

Comparison of the Selection Schemes

Fig. 7.10b shows the results with DARA-S and RSSI decisions for the 50/50 mix. Both are similar for the high-rate scenarios of 50 and 75 meters as these do not have to deal with low rates as result of the channel quality (cf. rate distribution of ideal adaptation in Fig. 7.2). With increasing scenario sizes, the probability of low-rate STAs increases. There, DARA-S outperforms RSSI decisions greatly: The number of VoIP STAs increases (with gains of 1, 3, and 3 VoIP STAs for 100 to 150 m scenarios). Note that the ideal rate distribution of the 150 m scenario (Fig. 7.2) is still far away from being a pure low-rate case as rates of 12 Mbps and below are applied in ten percent of the transmissions. Thus, this scenario represents the case of links with medium, fluctuating quality as can be found in typical indoor environments.

In these scenarios, DARA-S does also not significantly differ from the high *all Minstrel* curve in Fig. 7.10a, i.e., DARA-S leads to a similar performance as the pure Minstrel mix with the simple 'random in/out' decisions. We plotted the 'random in/out' selection curve of the *50/50 mix* in Fig. 7.10b. In all scenarios, the *50/50 mix* with 'random in/out' selection and RSSI decisions are not significantly different. In other words, applying RSSI does not lead to any improvements over simple random in/out decisions, where a STA becomes a handover candidate once its minimum QoS level is violated.

Accuracy Analysis of DARA-S

To evaluate the performance of our classifier recognizing low-rate STAs, we conducted an accuracy analysis [208] for DARA-S. For this, we applied an additional simulation

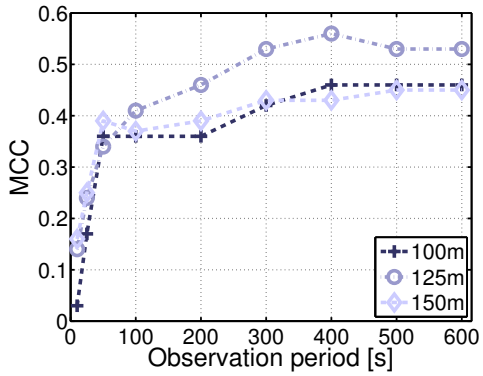


Figure 7.11: Matthews correlation coefficient for different Δt_{obs}

run for each edge length based on the 50/50 mix and observed the classification decisions plus the rate selection behavior afterwards for each node in each scenario. As a result, we can judge about the correctness of decisions included in *true positives (tp)* and *true negatives (tn)*, i.e., the classification is correct and STA's network access is granted (tp) or denied (tn) correctly. In contrast, *false positives (fp)* and *false negatives (fn)* cover wrong classifications, such that the STA is spuriously kept in the network (fp) or forced to conduct a handover (fn). For a single representation of these four cases, we selected *Matthews correlation coefficient (MCC)* as it is known to provide a "much more balanced evaluation" compared to an analysis, e.g., considering just simple true or false positive rates [208]. MCC is defined as

$$\text{MCC} = \frac{\text{tp}\,\text{tn} - \text{fp}\,\text{fn}}{\sqrt{(\text{tp}+\text{fn})(\text{tp}+\text{fp})(\text{tn}+\text{fp})(\text{tn}+\text{fn})}}. \tag{7.9}$$

MCC is the correlation coefficient between DARA-S's classifications and the observed rate selections of the STAs. A MCC of -1 represents a "total misalignment", +1 indicates a "total agreement", and 0 implies purely "random predictions" [208].

We calculated MCC for different durations Δt_{obs} ranging from 10 to 600 s over which we observed STA's rate selection process after the actual classification (Fig. 7.11). In the high-rate scenarios of 50 and 75 meters, MCC is not defined as these cases frequently lack true and false negatives [208]. For other scenarios with lower rates—and where as a result, also true and false negatives appear—MCC increases with longer observation periods. For Δt_{obs} of more than 50 s, MCC is above 0.3 in all scenarios demonstrating the classification accuracy of DARA-S even for short observation ranges of about a minute. Note that these results lie above simple randomized decisions (with MCCs around zero). The longer we observe after an estimation, the higher becomes the accuracy until it reaches its maximum.

7.8 Final Remarks: Considerations for Practical Usage

Regarding the 802.11 parameters discussed in Sec. 7.3.1, we now describe briefly, how they can be derived in our selected architecture for a given STA. For the signaling, we rely again on 802.11k/v.

Using the 802.11k "STA Statistics Request" (cf. Sec. 3.7.1), the AP triggers a STA to start measurements specified in this request. Specifically, the AP requests the STA to maintain a list of its transmission attempts which includes the fraction of successful and non-successful transmissions (i.e., without the reception of an ACK). After Δt_{trial}, the STA reports its data with the 802.11k "STA Statistics Response". In parallel, the AP keeps track of successfully received data frames and logs for each of them the data rate as well as the RSSI values. On the basis of these observations, the AP calculates the mean values and the standard deviations of the RSSI values, the number of transmission attempts per data frame, and the channel utilization. We propose to derive the mean number transmission attempts and its standard deviation following the approach given in Sec. 6.7.2. Further, on AP, we compute the pmf of the successfully received data frames. Regarding a support of the values for pmf, transmission attempts, and RSSI, we assume that the AP uses MIB variables relying on reserved fields of the "Transmit Stream/Category Measurement" for proprietary measures.

Finally, for the description of the device type triple, the AP additionally requests IEEE 802.11v "Diagnostic Reporting" information [6] (cf. Sec. 3.7.2) from the STA to obtain device information. Among others, the 802.11v diagnostic information includes the required triple consisting of a manufacturer identification, the NIC model, as well as its firmware version.

7.9 Conclusions

This chapter presented our novel data rate estimation scheme 'DARA' for WLAN networks. The core mechanism of DARA estimates the rate selection of a WLAN end-user device by observing its behavior on short time scales—without having any knowledge about the applied rate adaptation algorithm. We studied DARA's estimation accuracy for the selected RA schemes AARF/AMRR and Minstrel with data from simulations as well as from the WLAN testbed operated in our office environment. Our results show high accuracies for AARF/AMRR and certain estimation errors regarding the precise shape of the rate distribution with Minstrel. Nevertheless, the results indicate that DARA is indeed able to identify 'trends' of the rate selection for different RA schemes, i.e., whether a STA operates in the low or high-rate region. Finally, we utilized these trends in an application example, in which we used DARA's estimates as a basis for selecting suitable VoIP traffic in WLAN hotspots. Our performance evaluation between DARA and RSSI-based decisions shows significant gains in settings where a certain amount of end-user devices are struggling with fluctuating channels. Under these circumstances we gain up to 3 additional VoIP STAs, being about 8 percent, with DARA.

Chapter 8

Conclusions and Outlook

Under the umbrella of a joint framework tackling recent amendments from standardization as well as emerging trends from the perspective of cellular network operators, this thesis presents solutions for the preparation of resource-aware vertical handovers from the WLAN side, including improvements both on network as well as device level. Recent evolving directions make such vertical handovers a crucial issue for next generation wireless networks. The recent trend of multi-standard devices, usually always including means for WLAN access today, has become reality for a broad fraction of the world-wide human population with the smart phone and tablet wave. End users have become used to be connected anywhere and anytime to the Internet. As a result, they are interested in obtaining continuous services on application level with given perceived quality, thereby not primarily focussing on the amount of transported bytes in (wireless) networks that are required for a realization of such a service. This in turn has been shown to be a dominant challenge especially for cellular network operators, as on the one hand the amount of traffic to be served in their WWANs has been continuously growing as a result of an increasing number of end devices as well as new emerging applications such as video. On the other hand, a pure increase of the WWAN capacity, e.g., by recent technologies such as LTE or emerging innovations such as LTE-Advanced, is expected to be likely only one dimension of a solution. As revenues of operators less and less scale with the amount of transported traffic, the most promising, alternative dimension lies in the approach to offload traffic from WWANs to other, heterogeneous wireless access cells. Thereby, WLAN technology is seen as the most compelling offloading target as a result of its popularity and frequent availability at different spots ranging from home, to office, and cafe hotspot scenarios.

While today, WLANs are taken by end users usually whenever being available, either because of higher peak throughputs or due to a low-budget data transport, the next generation of offloading solutions discussed in 3GPP standardization bodies aims at 'seamless' shifts of traffic streams from WWANs to WLANs thus enabling a more fine-grained steering of the traffic on a flow level. In this direction, this thesis tackles that offloading as presented in 3GPP standardization bodies does not include the viewpoint of WLAN

owners or operators regarding a management of their radio resources. However, incorporating such a *resource-aware* perspective is in turn a key aspect when trying to maximize the number of end devices allocated to WLANs. For certain devices, regarding a resource management of WLAN hotspots, we propose to initiate handovers back to WWANs being denoted as *onloading*. Thereby, such backward handovers are supported already on end devices today for the case that a desired QoS level can not be supported in a WLAN.

8.1 Conclusions

Following the general motivation of resource-aware onloading handovers among different wireless access technologies, we make three major contributions throughout this thesis. In conclusion, we extended the support of WLANs towards resource-aware handovers to WWANs both on device as well as network level focussing on the following major research aspects.

On the device level, onloading handovers may suffer from the issue that a parallel support of both, on-going communication as well as a preparation of handovers may not be supported by the underlying radio hardware. In this direction, we present our innovative design for an *opportunistic preparation of handovers*, which enables a network discovery and alternative link setup together with a contiguous support of an on-going, strongly time-constrained communication pattern such as VoIP over a single transceiver chain. Being compliant with the existing IEEE 802.11 standard, we present different flavors of our design considering two distinct application cases. In the first, we study a neighbor network discovery of surrounding homogeneous WLAN cells for an IEEE 802.11 NIC inside an end device. First, with our numerical analysis we showed that our approach approximately doubles the maximum duration for a discovery of a given WLAN AP compared to the traditional passive scanning, while contrary being able to seamlessly support the transport of VoIP traffic at the same time. Next, with our simulative performance evaluation we showed that the costs of our scheme effectively lead to a reduction of the WLAN VoIP capacity by just one connection. In the second application case, motivated by the previous results, we show a standard-compliant approach enabling ongoing WLAN communication as well as a lengthy network entry in another heterogeneous technology using only a single, shared transceiver chain for both. Within our work, we considered in detail the network entry for WiMAX as it is known to last up to several seconds. Our analysis results did not only show the applicability of this approach over a wide range of parameters but further identify clear bounds on the WiMAX downlink load for which the WiFi communication does not suffer quality distortions.

Further, we consider to maximize the number of end devices allocated to WLANs, while at the same time aiming to uphold at least a minimum QoS level for each. As such, we show that it is beneficial to favor devices in WLAN operating in a resource-efficient way over devices that waste resources, e.g., as result of means for link layer adaptation or error recovery processes. Thus, we show the achievable gains if these 'unsuitable' devices are forced to conduct an onloading handover back to the WWANs. For a clear distinction of 'suitable' and 'unsuitable' devices, we present the design of

a performance metric allowing for an efficiency analysis regarding the occupied WLAN resources applicable either for all or a selected subset of devices. In a nutshell, the performance metric jointly combines available WLAN MAC and PHY measures to a unique criterion allowing a detailed insight with low computational effort and enabling a relative comparison of active WLAN devices. For a maximization of traffic served by a WLAN hotspot, we utilize the performance metric for the identification of onloading candidates in turn effectively enabling to accommodate other suitable devices in the WLAN cell. In our performance evaluation by means of simulations, we compared our scheme with the popular received signal strength and pure random decisions. In dense settings, where WLAN devices are capable to transmit at medium to high rates, our scheme outperforms the others strongly. The less devices are in regions with high rates, the more drop our gains, leading asymptotically to a similar behavior with classical received-signal-strength decisions. A second outcome of this thesis contribution is that it extends the results of sparse or dense device settings by further showing a tendency for operational points being of interest for wireless network operators. Dependent on the traffic mix, we propose different flavors of our selection scheme, thus effectively enabling a fine-tuning of the mix of devices being served by the WLAN hotspot.

Extending reactive decisions about the selection of handover candidates, the third contribution of this thesis tackles an estimation of the link layer behavior for a device that just has been associating with a WLAN cell. The major problem there lies in the issue that it is not known how this device will really behave. In real WLAN networks, varying vendor-specific behavior may make it impossible to judge the resource consumption of a device on the basis of the usual decision criteria such as received signal strength measures. This thesis does not only identify WLAN link data rate adaptation algorithms to be a major source of strong varying vendor specific behavior, but further presents the design of our link data rate estimation scheme 'DARA' that aims for a fast prediction of the data rate behavior for a given device on the basis of short observations. Our results highlight that our estimation scheme is capable to identify whether a device will rather fall into the low or the high data rate category, thus being able to support fine-grained reevaluation of handover decisions, in the best case even before a handover to WLAN was finally completed. Using the data rate estimations as criterion for decisions to which devices we enable continuous access to a WLAN cell and which ones are onloaded again, our performance evaluation showed significant gains under situations where certain devices perceive fluctuating channels.

In conclusion, in today's scenarios, where WLAN hotspots and WWANs are frequently managed and operated independently of each other, the contributions of this thesis are able to deliver a thorough support for WLAN owners and operators regarding a fine-grained tuning of their networks in high-load scenarios. Further, a direct *cooperation* between WWAN and WLAN operators will allow integrated approaches in the future considering a balanced handling of *both* handover directions for an off- and an onloading. For such emerging approaches, the contribution of this thesis will allow for a support of decisions regarding the joint capacity a heterogeneous network, thereby covering the aspects from the WLAN perspective.

8.2 Outlook

Yet, regarding our support of vertical handovers from the WLAN perspective, certain challenges still remain for future work. First, we see time-shifted, predictive traffic offloading to be not fully utilized in current approaches. Learning the end-user behavior—e.g., regarding his application in use on different mobile devices, his daily trajectory, but also his calendar entries and meeting schedules—may allow to estimate future traffic demands far before they actually appear. This may enable to stronger utilize wireless capacity by distributing the traffic smoothly in time, frequency, and space, jointly orchestrating traffic demands of end users as well as today's heterogeneous wireless accesses.

In a near future, cognitive radio approaches will make their way into 802.11 technology, allowing to utilize licensed portions of spectrum such as the TV Whitespace. However, these bands require arrangements to not harm the operation of the original owner of the spectrum. In the context of handovers from and to such WLANs, these requirements, e.g., regarding portions of usable spectrum and limits on transmit power, may make it even harder to evaluate the capacity of a network. Further, operation in such bands will affect handover mechanisms, as well-known techniques such as the active scanning in WLANs will be exposed to regulative limitations.

Given the significant gains achieved by our decision scheme considering WLAN for on-and offloading traffic, we expect future work to evaluate more complex systems dealing with the joint capacity of heterogeneous networks. To fully utilize these gains, a cooperation of WLAN owners and cellular operators is likely to pave the way towards a more efficient usage of wireless resources. With the increasing traffic load in WWANs as well as WLANs, we expect that operators will discover incentives to not only shift away traffic from their own networks but, in case of low loads, will also help out other networks by temporarily taken over a share of their load. We expect future approaches to aim at a cooperation and a *joint* coordination of load balancing mechanisms among heterogeneous networks. This may cover a broad range of topics, from architectures and protocols regarding an interworking of different networks and technologies, over different optimization criteria, to distributed mobility management approaches smoothly shifting traffic on a flow level.

Appendix A

Acronyms

3G	third generation of mobile telecommunications technology
4G	fourth generation of mobile telecommunications technology
3GPP	3rd Generation Partnership Project
AAA	authentication, authorization, and accounting
AARF	adaptive auto rate fallback
ABC	always best connected
AC	access category
AC_BE	AC for best effort traffic
AC_BK	AC for background traffic
AC_VI	AC for video traffic
AC_VO	AC for voice traffic
ACK	acknowledgment
aCWmin	minimum CW of a specific WLAN PHY
aCWmax	maximum CW of a specific WLAN PHY
AIFS	arbitration inter-frame space
AM	active mode
AMRR	adaptive multi rate retry
ANDI	access network discovery information

ANDSF	access network discovery and selection function
AP	access point
APC	AP controller
APSD	automatic power save delivery
ARD	automatic relevance determination
ARF	auto rate fallback
BC	backoff counter
BG STA	background-load STA
BS	base station
BSS	basic service set
BSSID	BSS identifier
CA	collision avoidance
CAPEX	capital expenditure
CAPWAP	Control and Provisioning of Wireless Access Points
CCA	clear channel assessment
CCK	complementary code keying
CDF	cumulative distribution function
CF-ACK	contention-free acknowledgment
CF-end	contention-free end
CF-poll	contention-free poll
CFP	contention-free period
CM	connection manager
CN	core network
CP	contention period
CRRM	common radio resource management
CSMA	carrier sense multiple access
CSMA/CA	CSMA with collision avoidance

CTS	clear-to-send
CW	contention window
CW_{min}	minimum contention window
CW_{max}	maximum contention window
DARA	data rate estimation
DARA-S	DARA-based selection
DBPSK	differential binary phase shift keying
DCD	downlink channel descriptor
DCF	distributed coordination function
DiffServ	differentiated services
DIFS	distributed inter-frame space
DL	downlink
DL-MAP	downlink map
DS	distribution system
DSCP	DiffServ code point
DSL	digital subscriber line
DSMIPv6	dual stack MIPv6
DSSS	direct sequence spread spectrum
EB	exabyte
EDCA	enhanced distributed channel access
EDCAF	EDCA function
ERP-OFDM	extended rate PHY OFDM
ESS	extended service set
ETSI	European Telecommunications Standards Institute
EW	equal weight
EWMA	exponentially weighted moving average
FCC	Federal Communications Commission

FCH	frame control header
FHSS	frequency hopping spread spectrum
FIFO	first-in-first-out
fn	false negatives
fp	false positives
FTP	File Transfer Protocol
FUSC	fully used sub-carrier
GAN	generic access network
GB	gigabyte
GHz	gigahertz
GP	Gaussian process
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GTP	GPRS Tunneling Protocol
GVR	goodput/VoIP-reduction ratio
HAL	hardware abstraction layer
HC	hybrid coordinator
HCCA	HCF controlled channel access
HCF	hybrid coordination function
HR/DSSS	high-rate DSSS
HSPA	High Speed Packet Access
HTTP	Hypertext Transfer Protocol
IAT	inter-arrival time
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IFOM	IP flow mobility
IP	Internet Protocol

ISM	industrial, scientific, and medical
ISMP	inter-system mobility policy
ISO	International Organization for Standardization
ISRP	inter-system routing policy
I-WLAN	interworking WLAN
LTE	Long Term Evolution
LWAPP	Lightweight Access Point Protocol
MAC	medium access control layer
MAPCON	multi-access packet data network connectivity
MB	megabyte
Mbps	megabits per second
MCC	Matthews correlation coefficient
MCS	modulation and coding scheme
MIB	management information base
MIH	media independent handover
MIHF	media independent handover function
MIMO	multiple input multiple output
MIP	mobile IP
ML	machine learning
MLME	MAC sublayer management entity
MPDU	MAC protocol data unit
MPTCP	Multipath TCP
MSD	multi-standard device
MSDU	MAC service data unit
MSE	mean squared error
MSGCF	MAC state generic convergence function
NAV	network allocation vector

NB-IFOM	network-based IFOM
NIC	network interface card
NRM	network reconfiguration manager
OFDM	orthogonal frequency division multiplexing
OPEX	operating expenditure
OS	operating system
OSI	Open Systems Interconnection
PC	point coordinator
PCF	point coordination function
PDN	packet data network
PDN GW	PDN gateway
PHY	physical layer
PIFS	PCF inter-frame space
PLC	packet loss concealment
PLCP	physical layer convergence procedure
PLME	physical layer management entity
PM	power management
pmf	probability mass function
PMIP	Proxy MIP
PS	power save
PSDU	PLCP service data unit
PTP	Precision Time Protocol
PUSC	partially use sub-carrier
QAM	quadrature amplitude modulation
QoE	quality of experience
QoS	quality of service
QPSK	quadrature phase shift keying

RA	rate adaptation
RAN	radio access network
RE	radio enabler
REQ	request
RMC	resource management controller
RRM	radio resource management
RSP	response
RSS	received signal strength
RSSI	received signal strength indicator
RTS	request-to-send
SAP	service access point
SAW	simple additive weighting
SCTP	Stream Control Transmission Protocol
SD	scan duration
SI	scan interval
SIFS	short inter-frame space
SINR	signal-to-interference and noise ratio
SME	station management entity
SNMP	Simple Network Management Protocol
SNR	signal-to-noise ratio
SP	service period
SRV	service field
STA	WLAN station
TB	terabyte
TBTT	target beacon transmission time
TCP	Transmission Control Protocol
TCP-BG	TCP background-load

TDD	time division duplex
THz	terahertz
TIM	traffic indication map
tn	true negatives
tp	true positives
TRM	terminal reconfiguration manager
TSF	timing synchronization function
TXOP	transmission opportunity
UCD	uplink channel descriptor
UL	uplink
UL-MAP	uplink map
UMTS	Universal Mobile Telecommunications System
UP	user priority
VoIP	voice over IP
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	wireless local area network
WWAN	wireless wide area network

Appendix B

The WLAN Simulation Model for ns-2

This chapter summarizes the simulation models being developed or applied throughout the work of this thesis within the network simulator ns-2. We start with a description and a verification of the TKN EDCA model which is publicly available on the Internet and frequently cited by other researchers [4, 209]. Afterwards we give a brief overview about the channel model consisting of a path loss and a fading component, before we discuss our extensions to the PHY model of ns-2. We note that parts of this chapter were published before in [4, 157].

B.1 TKN EDCA Model

The IEEE 802.11e task group finished the final draft version 13 in January 2005. The standardization process of the IEEE 802.11e task group lasted about five years, thus different modifications were incorporated from early draft versions to the final one. During the standardization process, various ns-2 models of 802.11e evolved, whereby we also developed a model for the predecessor of the EDCA [210]. Back in 2006, we updated our early model to include the final, standardized functionality of the EDCA to the, at that time, most recent version 2.28 of the network simulator [211].

While all the specifics about our model are detailed in [4], we briefly highlight the two basic design considerations in this section. Our model is an extension of the existing WLAN implementation included in ns-2 thus reusing the basic CSMA/CA feature. First, we introduced four interface queues between layer 3 and MAC. Dependent on their priority, packets are stored in one of these queues, whereby the 'prio' field in the IP header of each packet specifies the priority. Second, implementing one EDCAF instance per queue would have increased the complexity in the simulator since it may require a resolution entity for the internal collision handling. To keep things simple, we followed another approach instead: First, we have a single frame buffer for each AC at MAC. Second, we have variables storing contention parameters for each AC. Third, we have a single timing instance for the contention handling. With this instance, we schedule AIFS and a possible backoff for the AC that will expire first in time. In case that two or

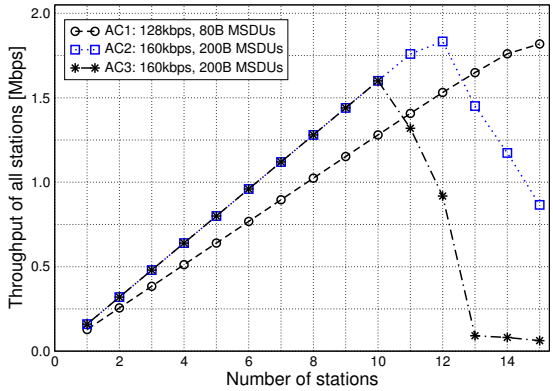


Figure B.1: Aggregated MAC layer throughput of all STAs

more priorities have the same smallest residual duration, the timer is scheduled for the highest priority. This approach is not only simple but also allows an easy resolution of internal collisions. Variables of lower ACs, such as residual backoff times, are adapted accordingly after certain events.

B.2 EDCA Model Verification

This section discusses results of the TKN EDCA model by showing its functionality with traffic of different ACs and its behavior when a wireless system changes from the non- to the saturated area. The considered scenario consists of a single AP with a different number of STAs, whereby all nodes are within the transmission range of each other. Each STA transmits three traffic streams in the uplink direction: a high-priority stream with 128 kbps and 80 Byte MSDUs, as well as a medium- and low-priority stream with 160 kbps and 200 Byte MSDUs, following the parameterization in Mangold et al. [212]. Although the authors use Poisson traffic for medium- and low-priority traffic, we rely on isochronous streams for the three ACs, since MAC's behavior will be investigated in the saturated area. Due to interface queues with lengths of 50 packets, Poisson arrivals average out in overload cases such that the MAC behavior dominates inter-transmission

Table B.1: Parameters for EDCA model verification, taken from [212]

	high	medium priority	low
AIFSN	2	4	7
CW_{min}	7	10	15
CW_{max}	7	31	255

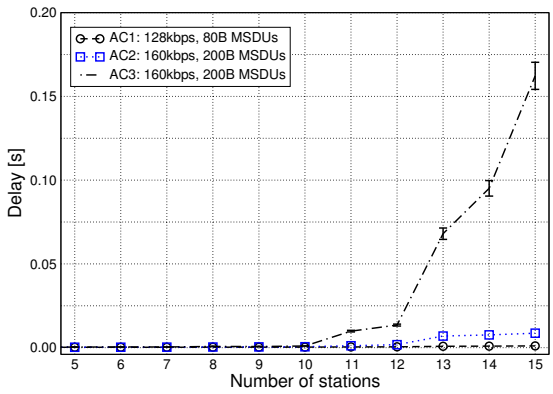


Figure B.2: Access delay for each AC

times. Table B.1 lists the contention parameters for the three ACs, which were also used in [210] already for the older model.

The first metric for the verification is the aggregated throughput per AC at MAC level of all STAs. Additionally, we measured the access delay for each AC which consists of the duration for the medium access contention, the (re)transmissions, as well as the ACK reception. For the evaluation of the results, we used the "Akaroa" tool [213] which enables to control the simulation duration as well as to conduct parallel simulations on different machines. The simulation process was terminated, when the mean value of a considered metric reached a confidence level of 95 percent, while staying within $+/-5$ percent around the mean.

With each simulation run, we increased the number of contending STAs by one, thereby expecting to see a degradation of the throughput and an increase in delay for the low-priority streams followed by the medium-priority traffic. The throughput curve in Fig. B.1 clearly shows this behavior when the scenario starts to enter the saturation case. There, the throughput of the lowest AC decreases for more than 10 STAs, while the throughput of the medium-priority streams degrades above 12 STAs. Far after both ACs suffer from strong degradations, the 128 *kbps* high-priority streams are not completely transmitted for more than 14 STAs. In line with these results are also the delay measures shown in Fig. B.2. Specifically the low-priority streams suffer from strong delays being effectively a result of the prioritized medium access. In conclusion, we see that the EDCA model is operating as expected, i.e., is capable of allowing a differentiated handling of stream-priorities at MAC level. Further, the interested reader is referred to our work in [4] that gives a comprehensive comparison of our current EDCA model version with previous implementations [210,212] covering also aspects from the standardization bodies before the finalization of the 802.11e amendment.

B.3 Channel Model

Radio signals, being essentially nothing else than electromagnetic waves, are usually exposed to different physical effects such as "reflection, diffraction, and scattering" as a result of the surrounding environment [214]. Due to movements of a sender or a receiver, as well as due to changes in the environment, the radio signal strength experienced at a receiver varies in the time and frequency domain. To account for these effects, we consider a channel model that consists of two parts: a *path loss* as well as *fading model*. With the first part, we consider the idealized, constant reduction of a wireless signal related to the distance between a sender and receiver, while the fading model brings the statistical properties of changes in the environment into our simulation environment.

We used the well-known *log-distance path loss model* to obtain the average receive power P_r at a receiver r with a distance d to the transmitter t. The reduction of the transmitted signal emitted with a transmit power P_t is known as the path loss PL and can be calculated as [214]

$$\overline{PL}(dB) = \overline{PL}(d_0) + 10\,n\,log\left(\frac{d}{d_0}\right), \tag{B.1}$$

with n, being the *path loss exponent* that is dependent on the selected radio environment, and the term $PL(d_0)$, representing the free space reference distance, commonly calculated for a distance from the transmitter of one meter [214]. Assuming receiver and antenna gains (G_t, G_r) as well as the system loss L to be one, $\overline{PL}(d_0)$ becomes

$$\begin{aligned} \overline{PL}(d_0) &= 10\,log\left\{\frac{P_t}{P_r}(d_0)\right\}, \\ \text{with } \frac{P_t}{P_r}(d_0) &= \frac{(4\pi d_0)^2 L}{G_t G_r \lambda^2} \approx \left(\frac{4\pi d_0}{\lambda}\right)^2, \end{aligned} \tag{B.2}$$

which in turn leads to a path loss (considering a reference distance d_0 of 1 m)

$$\overline{PL}(d) = 10\,log\left\{\left(\frac{4\pi}{\lambda}\right)^2 d^n\right\}. \tag{B.3}$$

Throughout this thesis, we have been considering large radio environments such as shopping malls, cafes or departure halls at an airport, thus having selected the path loss exponent n to be 2.8 [214, pp. 139–165].

Assuming slowest movements in the considered environment by pedestrians walking around at a speed of 2 km/h, let us quickly determine for which mean duration the wireless channel remains approximately unchanged. For this, the *coherence time* gives the time span "over which the channel impulse response is essentially invariant" [214]. The coherence time T_C is given by [214] as

$$T_C = \sqrt{\frac{9}{16\pi f_m^2}} = \frac{0.423}{f_m}, \tag{B.4}$$

Table B.2: Parameters of the PHY and Channel Model

Parameter	Notation	Value
transmit power	P_t	20 mW
center frequency	f_c	2.4 GHz
speed of light	c	$3 \cdot 10^8$ m/s
Boltzmann constant	k_b	$1.38 \cdot 10^{-23}$ J/K
channel bandwidth	B	20 MHz
temperature	T	290 K
Ricean k-factor	k	8.7
path-loss exponent	n	2.8

with f_m being the maximum Doppler shift

$$f_m = \frac{V}{\lambda} = \frac{V f_c}{c} \tag{B.5}$$

Having slowest movements at speed V of around 2 km/h = 0.556 m/s results in a coherence time (with center frequency f_c = 2.4 GHz, velocity of light c = 3 $10^8 m/s$) which is approximately equal to 95 ms.

In the considered radio environments, if one assumes to have a dominant path of the radio signal between transmitter and receiver, a *Ricean fading model* may be used for the description of the time-varying behavior. The name of the model derives from the *Ricean distribution*, which is "commonly used to describe the statistical time varying nature of the received envelope of a flat fading signal" [214] between transmitter and receiver. *Flat* thereby implies that the fading is not *frequency selective* [214]. Further, as calculated above, the coherence time is in our case around 95 ms, while a symbol for 802.11g lasts 4 μs [6]. Thus, we consider a *slow-fading channel* that varies much slower than a symbol duration [214].

To model such time-varying multi-path propagation effects, we used a Ricean fading model, whereby we rely on the implementation of Punnoose et al. [215] for the network simulator ns-2. For this model, the main parameter is the *Ricean K-Factor*, "which is defined as the ratio between the deterministic signal power and the variance of the multi path" [214]. We selected the K-Factor from a measurement campaign of Walker et al. [216], who conducted indoor measurements regarding the fading characteristics in a 2.4 GHz WLAN scenario, where changes in the environment are also influenced by people walking around. In our scenario aiming to model a radio environment such as a shopping mall, a cafe or a departure halls at an airport, we expect a similar influence by pedestrians.

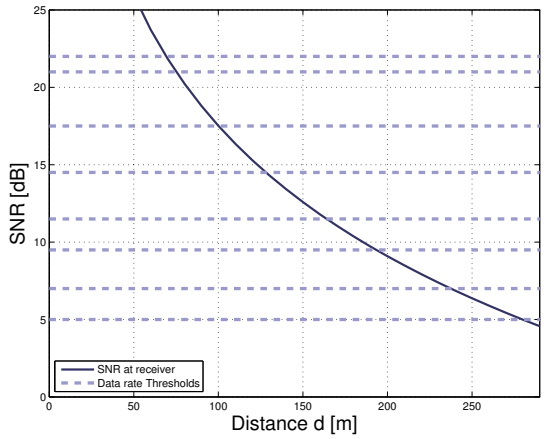


Figure B.3: SNR at the receiver in comparison to the data rate thresholds

B.4 PHY model

As discussed by Kochut et al. [217], the original wireless channel model for WLANs included in ns-2 does not accurately model *capture effects*. These effects occur when a strong wireless signal, e.g., of a sender being close to a receiver, superposes a weak signal. Now with the path loss and the Ricean fading model, capture effects may also occur in our simulations. Following the rationale of [217], we extended ns-2's wireless PHY model by an SINR part, where each receiver keeps track of the instantaneous power level present on the wireless channel. Then, for each single packet arriving at the receiver, we analyze the ratio of the signal level power to interference and noise disturbances, effectively deciding whether

- the PLCP preamble of the packet can be decoded correctly,
- the SINR is large enough to decode the PSDU (which may be transmitted at higher data rates).

Second, in case of multiple arriving packets at receiver's PHY, our SINR add-on is used to decide whether

- a frame is sufficiently stronger than others,
- an arriving frame is too weak to harm an already receiving one,
- a collision occurs between multiple frames.

For the different MCSs of IEEE 802.11g resulting into raw bit-rates ranging from 6 to 54 Mbps, we selected the different SNR thresholds as given by [218] for packet error rates of one percent, as shown in Table B.3. Note that the range of 5 to 22 dB for 6 to 54 Mbps also matches vendor-specific designs as detailed by [219].

Table B.3: SNR thresholds for the IEEE 802.11g link data rates

Data Rate [Mbps]	6	9	12	18	24	36	48	54
SNR [dB]	5.0	7.0	9.5	11.5	14.5	17.5	21.0	22.0

Finally, let us determine the different distances obtainable for each data rate SNR threshold, with the selected parameters of the PHY and the path loss model listed in Table B.2. For this we determine the SNR at the receiver r as

$$SNR_r = 10 \log \frac{P_r}{P_n}, \text{where } P_n = k_b \, T \, B. \tag{B.6}$$

Thereby, P_r is determined from Eq. B.3, thus leading to

$$SNR_r = 10 \log \frac{\lambda^2 P_t}{(4\pi)^2 P_n d^n}. \tag{B.7}$$

Fig. B.3 displays the SNR at a receiver dependent on the distance to a sender. Further, we visualize the selected SNR thresholds for the IEEE 802.11g data rates.

Appendix C

Auxiliary Means for the Thesis Preparation

The models and tools that we used during our research to obtain the results are mentioned and referenced in the corresponding chapters of this work. Further, this thesis was written and typeset with LaTeX and related packages from the Comprehensive TeX Archive Network (CTAN). The figures included in this work were generated by the software tools xmgrace, Adobe Illustrator 14.0, Microsoft Visio 2002 and 2007, and the plotting functions of Matlab. In the Visio plots, we partially relied on predefined shapes included in the software package.

Appendix D

List of Publications

In the following, we give the list of publications that have been compiled throughout the work of this thesis, grouped into journal papers and book chapters, conference proceedings, patents, and TKN technical reports.

D.1 Journal Papers and Book Chapters

Sven Wiethölter and Marc Emmelmann, Chapter: "Modeling handover from the access networks perspective", Modeling and Tools for Network Simulation, Klaus Wehrle, Mesut Günes and James Gross (Editors), Springer, 2010.

Murad Abusubaih, Sven Wiethölter, James Gross, and Adam Wolisz, "A new access point selection policy for multi-rate IEEE 802.11 WLANs," International Journal of Parallel, Emergent and Distributed Systems (IJPEDS), vol. 23, no. 4, pp. 291–307, 2008.

Marc Emmelmann, Sven Wiethölter, Andreas Köpsel, Cornelia Kappler, and Adam Wolisz, "Moving towards seamless mobility – state of the art and emerging aspects in standardization bodies," Springer's International Journal on Wireless Personal Communication – Special Issue on Seamless Handover in Next Generation Wireless / Mobile Networks, vol. 43, no. 3, pp. 803–816, 2007.

D.2 Conference Proceedings

Sven Wiethölter, Andreas Ruttor, Uwe Bergemann, Manfred Opper, and Adam Wolisz, "DARA: estimating the behavior of data rate adaptation algorithms in WLAN hotspots," in Proc. of the 32nd IEEE International Conference on Computer Communications (INFOCOM '13), Mini-Conference Track, pp. 280–284, Apr. 2013.

Sven Wiethölter, Marc Emmelmann, Robert Andersson, and Adam Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in Proc. of the IEEE International Conference on Communications (ICC'12), pp. 5423–5428, Jun. 2012.

Sven Wiethölter, Marc Emmelmann, Yerong Chen, and Adam Wolisz, "On the analysis of WiFi communication and WiMAX network entry over single radios," in Proc. of the second IEEE Workshop on Convergence among Heterogeneous Wireless Systems in Future Internet (ICC'12 WS – CONWIRE), pp. 5665–5669, Jun. 2012.

Marc Emmelmann, Sven Wiethölter, and Hyung-Taek Lim, "Influence of network load on the performance of opportunistic scanning," in Proc. of the IEEE Conference on Local Computer Networks (LCN '09), pp. 1–8, Oct. 2009.

Marc Emmelmann, Sven Wiethölter, and Hyung-Taek Lim, "Opportunistic scanning: interruption-free network topology discovery for wireless mesh networks," in Proc. of the 10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '09), pp. 1–6, Jun. 2009.

Sven Wiethölter and Adam Wolisz, "Selecting vertical handover candidates in IEEE 802.11 mesh networks," in Proc. of the first IEEE WoWMoM Workshop on Hot Topics in Mesh Networking (HotMESH), pp. 1–7, Jun. 2009.

Marc Emmelmann, Sven Wiethölter, Andreas Köpsel, Cornelia Kappler, and Adam Wolisz, "Moving towards seamless mobility: state of the art and emerging aspects in standardization bodies," in Proc. of the 9th International Symposium on Wireless Personal Multimedia Communications (WPMC '06), Sep. 2006, invited paper.

Murad Abusubaih, James Gross, Sven Wiethölter, and Adam Wolisz, "On access point selection in IEEE 802.11 wireless local area networks," in Proc. of the Sixth International Workshop on Wireless Local Networks (WLN '06), Nov. 2006.

D.3 Patents

Marc Emmelmann, Sven Wiethölter, and Hyung-Taek Lim, "Network discovery," pending patent application, USPTO US 61/186,528, Nov. 2011.

D.4 TKN Technical Reports

Sven Wiethölter, Andreas Ruttor, Uwe Bergemann, Manfred Opper, and Adam Wolisz, "DARA: estimating the behavior of data rate adaptation algorithms in WLAN hotspots," Telecommunication Networks Group, Technische Universität Berlin, TKN Technical Report Series, TKN-13-001, Jan. 2013.

Sven Wiethölter and Adam Wolisz, "Selecting vertical handover candidates in WLAN hotspots," Telecommunication Networks Group, Technische Universität Berlin, TKN Technical Report Series, TKN-08-009, Aug. 2008.

Sven Wiethölter, Marc Emmelmann, Christian Hoene, and Adam Wolisz, "TKN EDCA model for ns-2," Telecommunication Networks Group, Technische Universität Berlin, TKN Technical Report Series, TKN-06-003, Jun. 2006.

Sven Wiethölter, "Virtual utilization and VoIP capacity of WLANs supporting a mix of data rates," Telecommunication Networks Group, Technische Universität Berlin, TKN Technical Report Series, TKN-05-004, Sep. 2005.

Bibliography

[1] J. Schiller, *Mobile Communications*. Addison-Wesley, 2000.

[2] G. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. Costa, and B. Walke, "The IEEE 802.11 universe," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 62–70, Jan. 2010.

[3] B. O'Hara and A. Petrick, *IEEE 802.11 Handbook: A Designer's Companion*, 2nd ed., ser. IEEE Standards Wireless Networks Series. New York: IEEE Press, 2005.

[4] S. Wiethölter, M. Emmelmann, C. Hoene, and A. Wolisz, "TKN EDCA model for ns-2," TKN, TU Berlin, TKN Technical Report Series TKN-06-003, Jun. 2006.

[5] M. Emmelmann, S. Wiethölter, and H.-T. Lim, "Opportunistic scanning: Interruption-free network topology discovery for wireless mesh networks," in *Proc. of the 10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '09)*, Jun. 2009, pp. 1–6.

[6] *IEEE Std 802.11-2012—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.11-2012, Mar. 2012, Revision of IEEE Std. 802.11-2007.

[7] A. S. Tanenbaum, *Computer Networks*, 3rd ed. Prentice Hall, 1996.

[8] M. Abusubaih, S. Wiethölter, J. Gross, and A. Wolisz, "A new access point selection policy for multi-rate IEEE 802.11 WLANs," *International Journal of Parallel, Emergent and Distributed Systems (IJPEDS)*, vol. 23, no. 4, pp. 291–307, Aug. 2008.

[9] A. Mishra, N. L. Petroni, Jr., W. A. Arbaugh, and T. Fraser, "Security issues in IEEE 802.11 wireless local area networks: a survey," *Wireless Communications and Mobile Computing – Special Issue: Emerging WLAN Applications and Technologies*, vol. 4, no. 8, pp. 821–833, Dec. 2004.

[10] J.-C. Chen, M.-C. Jiang, and Y.-W. Liu, "Wireless LAN security and IEEE 802.11i," *IEEE Wireless Communications*, vol. 12, no. 1, pp. 27–36, Feb. 2005.

[11] *802.11i-2004—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 6: Medium Access Control (MAC) Security Enhancements*, IEEE Std 802.11i-2004, Jul. 2004, Amendment to IEEE Std 802.11-1999 (Reaff 2003).

[12] *IEEE Std 802.11-1997—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.11-1997, Jun. 1997.

[13] *IEEE Std 802.11b-1999—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band*, IEEE Std 802.11b-1999, Sep. 1999, Supplement to IEEE Std 802.11-1999.

[14] *IEEE Std 802.11a-1999—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band*, IEEE Std 802.11a-1999, 1999, Supplement to IEEE Std 802.11-1999.

[15] *IEEE Std 802.11g-2003—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 4: Further Higher Data Rate Extension in the 2.4 GHz Band*, IEEE Std 802.11g-2003, Jun. 2003, Amendment to IEEE Std 802.11-1999 (Reaff 2003).

[16] *IEEE Std 802.11n-2009—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 5: Enhancements for Higher Throughput*, IEEE Std 802.11n-2009, Oct. 2009, Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009.

[17] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "802.11 with multiple antennas for dummies," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 19–25, Jan. 2010.

[18] D. Vassis, G. Kormentzas, A. Rouskas, and I. Maglogiannis, "The IEEE 802.11g standard for high data rate WLANs," *IEEE Network*, vol. 19, no. 3, pp. 21–26, 2005.

[19] Q. Ni, L. Romdhani, and T. Turletti, "A survey of QoS enhancements for IEEE 802.11 wireless LAN," *Journal of Wireless Communications and Mobile Computing*, vol. 4, no. 5, pp. 547–566, Aug. 2004.

[20] *IEEE Std 802.11e-2005—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 8: Medium Access Control (MAC) Quality of Service (QoS) Enhancements*, IEEE Std 802.11e-2005, Sep. 2005, Amendment to IEEE Std. 802.11-1999 (Reaff 2003).

[21] Q. Ni, "Performance analysis and enhancements for IEEE 802.11e wireless networks," *IEEE Network*, vol. 19, no. 4, pp. 21–27, 2005.

[22] N. Ramos, D. Panigrahi, and S. Dey, "Quality of service provisioning in 802.11e networks: challenges, approaches, and future directions," *IEEE Network*, vol. 19, no. 4, pp. 14–20, 2005.

[23] D. Skyrianoglou, N. I. Passas, and A. K. Salkintzis, "ARROW: an efficient traffic scheduling algorithm for IEEE 802.11e HCCA," *IEEE Transactions on Wireless Communications*, vol. 5, no. 12, pp. 3558–3567, 2006.

[24] A. L. Ruscelli, G. Cecchetti, A. Mastropaolo, and G. Lipari, "A greedy reclaiming scheduler for IEEE 802.11e HCCA real-time networks," in *Proc. of the 14th ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '11)*, 2011, pp. 223–230.

[25] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 headers," RFC 2474 (Proposed Standard), Internet Engineering Task Force, Dec. 1998, updated by RFCs 3168, 3260.

[26] *IEEE Std 802.1D-2004—Media Access Control (MAC) Bridges*, IEEE Std 802.1D-2004, Feb. 2004, Revision of IEEE Std 802.1D-1998.

[27] M. Emmelmann, S. Wiethölter, A. Köpsel, C. Kappler, and A. Wolisz, "Moving towards seamless mobility: State of the art and emerging aspects in standardization bodies," in *Proc. of the 9th International Symposium on Wireless Personal Multimedia Communications (WPMC '06)*, Sep. 2006, invited Paper.

[28] ——, "Moving towards seamless mobility: State of the art and emerging aspects in standardization bodies," *Springer's International Journal on Wireless Personal Communication — Special Issue on Seamless Handover in Next Generation Wireless/Mobile Networks*, vol. 43, no. 3, pp. 803–816, 2007.

[29] S. Wiethölter and M. Emmelmann, *Modeling and Tools for Network Simulation*. Springer, Mar. 2010, ch. Modeling Handover from the Access Networks Perspective, pp. 341–356.

[30] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2008 – 2013," Whitepaper, Jun. 2009.

[31] ——, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009 – 2014," Whitepaper, Feb. 2010.

[32] ——, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 – 2015," Whitepaper, Feb. 2011.

[33] ——, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011 – 2016," Whitepaper, Feb. 2012.

[34] ——, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012 – 2017," Whitepaper, Feb. 2013.

[35] T. Farrar, "Is Cisco stacking the deck with its mobile data numbers?" Feb. 2013. [Online]. Available: http://gigaom.com/2013/02/09/is-cisco-stacking-the-deck-with-its-mobile-data-numbers/

[36] D. Bubley, "New Cisco mobile VNI numbers bring some realism," Feb. 2013. [Online]. Available: http://disruptivewireless.blogspot.de/2013/02/new-cisco-mobile-vni-numbers-bring-some.html

[37] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. of the 30th IEEE International Conference on Computer Communications (INFOCOM '11)*, 2011, pp. 882–890.

[38] Y. Jin, N. Duffield, A. Gerber, P. Haffner, W.-L. Hsu, G. Jacobson, S. Sen, S. Venkataraman, and Z.-L. Zhang, "Characterizing data usage patterns in a large cellular network," in *Proc. of the ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design (CellNet '12)*, 2012, pp. 7–12.

[39] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*, 2010, pp. 209–222.

[40] K. Lee, I. Rhee, J. Lee, S. Chong, and Y. Yi, "Mobile data offloading: how much can WiFi deliver?" in *Proc. of the ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT '10)*, 2010, pp. 1–12.

[41] S. Liu and A. Striegel, "Casting doubts on the viability of WiFi offloading," in *Proc. of the ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design (CellNet '12)*, 2012, pp. 25–30.

[42] Z. Jiang and L. Kleinrock, "Web prefetching in a mobile environment," *IEEE Personal Communications*, vol. 5, no. 5, pp. 25–34, 1998.

[43] N. Ristanovic, J.-Y. Le Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *Proc. of the IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS '11)*, 2011, pp. 202–211.

[44] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: a case study," in *Proc. of the 5th ACM Workshop on Challenged Networks (CHANTS '10)*. ACM, 2010, pp. 31–38.

[45] P. Baier, F. Dürr, and K. Rothermel, "TOMP: Opportunistic traffic offloading using movement predictions," in *Proc. of the IEEE Conference on Local Computer Networks (LCN '12)*. IEEE, 2012, pp. 50–58.

[46] J. Manner and M. Kojo, "Mobility related terminology," RFC 3753 (Informational), Internet Engineering Task Force, Jun. 2004.

[47] I. Akyildiz, J. Xie, and S. Mohanty, "A survey of mobility management in next-generation all-IP-based wireless systems," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 16–28, 2004.

[48] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Communications*, vol. 10, no. 1, pp. 49–55, 2003.

[49] G. Wu, Q. Li, R. Q. Hu, and Y. Qian, *Heterogeneous Cellular Networks*. John Wiley & Sons Inc., 2013, ch. 1: Overview of Heterogeneous Networks, pp. 1–25.

[50] A. Doufexi, E. Tameh, A. Nix, S. Armour, and A. Molina, "Hotspot wireless LANs to enhance the performance of 3G and beyond cellular networks," *IEEE Communications Magazine*, vol. 41, no. 7, pp. 58–65, 2003.

[51] W. Wang, X. Liu, J. Vicente, and P. Mohapatra, "Integration gain of heterogeneous WiFi/WiMAX networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 8, pp. 1131–1143, 2011.

[52] M. D. Nisar, V. Pauli, and E. Seidel, "Multi-RAT traffic steering – Why, when, and how could it be beneficial?" Whitepaper, Dec. 2011, Nomor Research.

[53] P. Munoz, R. Barco, D. Laselva, and P. Mogensen, "Mobility-based strategies for traffic steering in heterogeneous networks," *IEEE Communications Magazine*, pp. 54–62, May 2013.

[54] S. J. Lincke, "Vertical handover policies for common radio resource management," *International Journal of Communication Systems*, vol. 18, no. 6, pp. 527–543, 2005.

[55] A. Tolli, P. Hakalin, and H. Holma, "Performance evaluation of common radio resource management (CRRM)," in *Proc. of the IEEE International Conference on Communications (ICC '02)*, vol. 5, 2002, pp. 3429–3433.

[56] A.-E. M. Taha, H. S. Hassanein, and H. T. Mouftah, "Vertical handoffs as a radio resource management tool," *Computer Communications*, vol. 31, no. 5, pp. 950–961, 2008.

[57] J. Márquez-Barja, C. M. T. Calafate, J.-C. Cano, and P. Manzoni, "An overview of vertical handover techniques: Algorithms, protocols and tools," *Computer Communications*, vol. 34, no. 8, pp. 985–997, 2011.

[58] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Computer Communications*, vol. 31, no. 10, pp. 2607–2620, Jun. 2008.

[59] X. Yan, Y. A. Sekercioglu, and S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks," *Computer Networks*, vol. 54, no. 11, pp. 1848–1863, 2010.

[60] L. Wang and G.-S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks – a tutorial," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 271–292, 2013.

[61] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 104–112, 2013.

[62] A. Ghosal, "Mobile data offload: Can Wi-Fi deliver?" Whitepaper, Jan. 2010.

[63] M. Olsson, S. Sultana, S. Rommer, L. Frid, and C. Mulligan, *EPC and 4G Packet Networks: Driving the Mobile Broadband Revolution*, 2nd ed. Academic Press, 2013.

[64] M. Wasserman and P. Seite, "Current practices for multiple-interface hosts," RFC 6419 (Informational), Internet Engineering Task Force, Nov. 2011.

[65] Google Inc., "Android Developers: ConnectivityManager." [Online]. Available: http://developer.android.com/reference/android/net/ConnectivityManager.html

[66] K. Bell, "iOS 6 allows apps to use 'Wi-Fi Plus Cellular' connections," Aug. 2012. [Online]. Available: http://www.cultofmac.com/183414/ios-6-allows-apps-to-use-wi-fi-plus-cellular-connections/

[67] C. Zibreg, "Where's my iOS Wi-Fi Plus Cellular toggle?" Sep. 2012. [Online]. Available: http://www.idownloadblog.com/2012/09/25/wi-fi-plus-cellular-missing-in-ios-6/

[68] J. Smith, "iOS 7 Beta 3 brings back smart WiFi feature & more," Jul. 2013. [Online]. Available: http://www.gottabemobile.com/2013/07/08/ios-7-beta-3-brings-back-smart-wifi-feature-more/

[69] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "Tcp extensions for multipath operation with multiple addresses," RFC 6824 (Experimental), Jan. 2013.

[70] L. Ong and J. Yoakum, "An Introduction to the Stream Control Transmission Protocol (SCTP)," RFC 3286 (Informational), May 2002.

[71] 3GPP, "3GPP system to Wireless Local Area Network (WLAN) interworking; System description (Release 6)," TS 23.234 V6.10.0, Sep. 2006.

[72] ——, "Generic Access Network (GAN); Stage 2 (Release 6)," TS 43.318 V6.12.0, May 2008.

[73] R. Ferrus, O. Sallent, and R. Agustí, "Interworking in heterogeneous wireless networks: Comprehensive framework and future trends," *IEEE Wireless Communications*, vol. 17, no. 2, pp. 22–31, 2010.

[74] K. Andersson, "Interworking Techniques and Architectures for Heterogeneous Wireless Networks," *Journal of Internet Services and Information Security (JISIS)*, vol. 2, no. 1, pp. 22–48, Feb. 2012.

[75] D.-E. Meddour, U. Javaid, N. Bihannic, T. Rasheed, and R. Boutaba, "Completing the convergence puzzle: a survey and a roadmap," *IEEE Wireless Communications*, vol. 16, no. 3, pp. 86–96, 2009.

[76] C. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 46–53, 2012.

[77] 3GPP, "IP flow mobility and seamless Wireless Local Area Network (WLAN) offload; Stage 2 (Release 10)," TS 23.261 V10.2.0, Mar. 2012.

[78] A. de la Oliva, C. J. Bernardos, M. Calderon, T. Melia, and J. C. Zuniga, "IP flow mobility: smart traffic offload for future wireless networks," *IEEE Communications Magazine*, vol. 49, no. 10, pp. 124–132, 2011.

[79] J. Korhonen, T. Savolainen, A. Ding, and M. Kojo, "Toward network controlled IP traffic offloading," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 96–102, 2013.

[80] J. C. Zuniga, C. J. Bernardos, A. de la Oliva, T. Melia, R. Costa, and A. Reznik, "Distributed mobility management: a standards landscape," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 80–87, 2013.

[81] N. Dimitriou, L. Sarakis, D. Loukatos, G. Kormentzas, and C. Skianis, "Vertical handover (VHO) framework for future collaborative wireless networks," *International Journal of Network Management*, vol. 21, no. 6, pp. 548–564, 2011.

[82] 3GPP, "Architecture enhancements for non-3GPP accesses (Release 11)," TS 23.402 V11.6.0, Mar. 2013.

[83] *IEEE Std 1900.4-2009—IEEE Standard for Architectural Building Blocks Enabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks*, IEEE Std 1900.4-2009, Feb. 2009.

[84] S. Buljore, H. Harada, S. Filin, P. Houze, K. Tsagkaris, O. Holland, K. Nolte, T. Farnham, and V. Ivanov, "Architecture and enablers for optimized radio resource usage in heterogeneous wireless access networks: The IEEE 1900.4 Working Group," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 122–129, 2009.

[85] F. Granelli, P. Pawelczak, R. Prasad, K. P. Subbalakshmi, R. Chandramouli, J. Hoffmeyer, and H. Berger, "Standardization and research in cognitive and dynamic spectrum access networks: IEEE SCC41 efforts and other activities," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 71–79, 2010.

[86] H. Harada, Y. Alemseged, S. Filin, M. Riegel, M. Gundlach, O. Holland, B. Bochow, M. Ariyoshi, and L. Grande, "IEEE dynamic spectrum access networks standards committee," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 104–111, 2013.

[87] *IEEE Std 802.21-2008—Media Independent Handover Services*, IEEE Std 802.21-2008, Jan. 2009.

[88] J. Sachs, "A generic link layer for future generation wireless networking," in *Proc. of IEEE International Conference on Communications (ICC '03)*, vol. 2, 2003, pp. 834–838.

[89] B. S. Ghahfarokhi and N. Movahhedinia, "A survey on applications of IEEE 802.21 Media Independent Handover framework in next generation wireless networks," *Computer Communications*, vol. 36, no. 10–11, pp. 1101–1119, Jun. 2013.

[90] R. Silva, P. Carvalho, P. Sousa, and P. Neves, "Enabling Heterogeneous Mobility in Android Devices," *Mobile Networks and Applications*, vol. 16, no. 4, pp. 518–528, Aug. 2011.

[91] *IEEE Std 802.11-2007—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.11-2007, Jun. 2007, Revision of IEEE Std. 802.11-1999.

[92] *IEEE 802.11k-2008—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 1: Radio Resource Measurement of Wireless LANs*, IEEE Std 802.11k-2008, Jun. 2008, Amendment to IEEE Std 802.11-2007.

[93] *IEEE 802.11v—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Amendment 8: IEEE 802.11 Wireless Network Management*, IEEE Std 802.11v-2011, Sep. 2011, amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, IEEE Std 802.11w-2009, IEEE Std 802.11n-2009, IEEE Std 802.11p-2010, and IEEE Std 802.11z-2010.

[94] R. Murty, J. Padhye, R. Chandra, A. Wolman, and B. Zill, "Designing high performance enterprise Wi-Fi networks," in *Proc. of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI '08)*, 2008, pp. 73–88.

[95] M. Abusubaih, J. Gross, S. Wiethölter, and A. Wolisz, "On access point selection in IEEE 802.11 wireless local area networks," in *Proc. of the 31st IEEE Conference on Local Computer Networks (LCN '06)*, Nov. 2006, pp. 879–886.

[96] C. Rossi, C. Casetti, and C. Chiasserini, "Bandwidth monitoring in multi-rate 802.11 WLANs with elastic traffic awareness," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM '11)*, Dec. 2011, pp. 1–5.

[97] M. Corici, J. Fiedler, T. Magedanz, and D. Vingarzan, "Access Network Discovery and Selection in the Future Wireless Communication," *Mobile Networks and Applications*, vol. 16, no. 3, pp. 337–349, Jun. 2011.

[98] N. Vallina-Rodriguez, V. Erramilli, Y. Grunenberger, L. Gyarmati, N. Laoutaris, R. Stanojevic, and K. Papagiannaki, "When David helps Goliath: the case for 3G onloading," in *Proc. of the 11th ACM Workshop on Hot Topics in Networks (Hotnets '12)*, 2012, pp. 85–90.

[99] C. Rossi, N. Vallina-Rodriguez, V. Erramilli, Y. Grunenberger, L. Gyarmati, N. Laoutaris, R. Stanojevic, K. Papagiannaki, and P. Rodriguez, "3GOL: power-boosting ADSL using 3G onloading," in *Proc. of the 9th ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT '13)*, 2013, pp. 187–198.

[100] J. Rabaey, A. Wolisz, A. Ercan, A. Araujo, F. Burghardt, S. Mustafa, A. Parsa, S. Pollin, I.-H. Wang, and P. Malagon, "Connectivity Brokerage - enabling seamless cooperation in wireless networks (a white paper)," Oct. 2010, whitepaper.

[101] X. Yang, D. Akhmetov, and H.-Y. Liu, "RF concurrency performance study for integrated WiFi radio," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM '10)*, 2010, pp. 1–5.

[102] D. Gao, J. Cai, and K. N. Ngan, "Admission control in IEEE 802.11e wireless LANs," *IEEE Network*, vol. 19, no. 4, pp. 6–13, 2005.

[103] L. Khoukhi, H. Badis, L. Merghem-Boulahia, and M. Esseghir, "Admission control in wireless ad hoc networks: a survey," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, 2013.

[104] P. Raptis, V. Vitsas, and K. Paparrizos, "Packet delay metrics for IEEE 802.11 Distributed Coordination Function," *Mobile Networks and Applications*, vol. 14, no. 6, pp. 772–781, Dec. 2009.

[105] B. Hamdaoui, M. Elaoud, and P. Ramanathan, "A delay-based admission control mechanism for multimedia support in IEEE 802.11e wireless LANs," *Wireless Networks*, vol. 15, no. 7, pp. 875–886, Oct. 2009.

[106] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *Proc. of the 22nd IEEE International Conference on Computer Communications (INFOCOM '03)*, vol. 2, 2003, pp. 836–843.

[107] B. Augustin and A. Mellouk, "On traffic patterns of HTTP applications," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM '11)*, 2011, pp. 1–6.

[108] A. Dainotti, A. Pescape, and K. Claffy, "Issues and future directions in traffic classification," *IEEE Network*, vol. 26, no. 1, pp. 35–40, 2012.

[109] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[110] G. Maier, F. Schneider, and A. Feldmann, "A first look at mobile hand-held device traffic," in *Proc. of the 11th International Conference on Passive and Active Measurement (PAM '10), Springer*, 2010, pp. 161–170.

[111] Nokia Siemens Networks, "Improving 4G coverage and capacity indoors and at hotspots with LTE femtocells," Whitepaper, 2011.

[112] Skype Communications SARL, "FAQ: Does Skype for iPad work over 3G, 4G or WiFi?" [Online]. Available: http://support.skype.com/en/faq/FA12237

[113] Viber Media Inc., "How does VoIP and Viber for iPhone work?" [Online]. Available: http://helpme.viber.com/Knowledgebase/Article/View/91/6/how-does-voip-and-viber-for-iphone-work

[114] Kineto Wireless, "Smart offload for smartphones," Whitepaper, 2010.

[115] A. Gember, A. Anand, and A. Akella, "A comparative study of handheld and non-handheld traffic in campus Wi-Fi networks," in *Proc. of the 12th International Conference on Passive and Active Measurement (PAM '11), Springer*, 2011, pp. 173–183.

[116] J. Case, M. Fedor, M. Schoffstall, and J. Davin, "Simple Network Management Protocol (SNMP)," RFC 1157 (Historic), Internet Engineering Task Force, May 1990.

[117] InterDigital, "Cellular-Wi-Fi integration: A comprehensive analysis of the technology and standardization roadmap," Jun. 2012, whitepaper.

[118] P. Calhoun, M. Montemurro, and D. Stanley, "Control and Provisioning of Wireless Access Points (CAPWAP) protocol binding for IEEE 802.11," RFC 5416 (Proposed Standard), Internet Engineering Task Force, Mar. 2009.

[119] P. Calhoun, R. Suri, N. C. Winget, M. Williams, S. Hares, B. O'Hara, and S. Kelly, "Lightweight Access Point Protocol," RFC 5412 (Historic), Internet Engineering Task Force, Feb. 2010.

[120] ITU-T Recommendation G.711, "Pulse code modulation (PCM) of voice frequencies," Nov. 1998.

[121] M. Emmelmann, S. Wiethölter, and H.-T. Lim, "Influence of network load on the performance of opportunistic scanning," in *Proc. of the IEEE Conference on Local Computer Networks (LCN '09)*, Oct. 2009, pp. 1–8.

[122] S. Wiethölter, M. Emmelmann, Y. Chen, and A. M. Wolisz, "On the analysis of WiFi communication and WiMAX network entry over single radios," in *Proc. of the second IEEE Workshop on Convergence among Heterogeneous Wireless Systems in Future Internet (ICC'12 WS - CONWIRE)*, Jun. 2012, pp. 5665–5669.

[123] A. Mishra, M. Shin, and W. Arbaugh, "An empirical analysis of the IEEE 802.11 MAC layer handoff process," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, pp. 93–102, Apr. 2003.

[124] I. Ramani and S. Savage, "Syncscan: practical fast handoff for 802.11 infrastructure networks," in *Proc. of the 24th IEEE International Conference on Computer Communications (INFOCOM '05)*, vol. 1, no. 1, Mar. 2005, pp. 675–684.

[125] Y.-S. Chen, M.-C. Chuang, and C.-K. Chen, "DeuceScan: deuce-based fast handoff scheme in IEEE 802.11 wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1126–1141, 2008.

[126] H. Velayos and G. Karlsson, "Techniques to reduce the IEEE 802.11b handoff time," in *Proc. of the IEEE International Conference on Communications (ICC '04)*, vol. 7, 2004, pp. 3844–3848.

[127] H. Wu, K. Tan, Y. Zhang, and Q. Zhang, "Proactive Scan: fast handoff with smart triggers for 802.11 Wireless LAN," in *Proc. of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, 2007, pp. 749–757.

[128] J. Teng, C. Xu, W. Jia, and D. Xuan, "D-Scan: enabling fast and smooth handoffs in AP-dense 802.11 wireless networks," in *Proc. of the 28th IEEE International Conference on Computer Communications (INFOCOM '09)*, 2009, pp. 2616–2620.

[129] S. Pack, J. Choi, T. Kwon, and Y. Choi, "Fast-handoff support in IEEE 802.11 wireless networks," *IEEE Communications Surveys and Tutorials*, vol. 9, no. 1, pp. 2–12, 2007.

[130] Y. Liao and L. Cao, "Practical schemes for smooth MAC layer handoff in 802.11 wireless networks," in *Proc. of the 7th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '06)*, 2006, pp. 181–190.

[131] R. Chandra, P. Bahl, and P. Bahl, "MultiNet: connecting to multiple IEEE 802.11 networks using a single wireless card," in *Proc. of the 23rd IEEE International Conference on Computer Communications (INFOCOM '04)*, vol. 2, 2004, pp. 882–893.

[132] Y. Chen, N. Smavatkul, and S. Emeott, "Power management for VoIP over IEEE 802.11 WLAN," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC '04)*, vol. 3, 2004, pp. 1648–1653.

[133] A. Nicholson, S. Wolchok, and B. Noble, "Juggler: Virtual networks for fun and profit," *IEEE Transactions on Mobile Computing*, vol. 9, no. 1, pp. 31–43, 2010.

[134] D. Giustiniano, E. Goma, A. Lopez, and P. Rodriguez, "WiSwitcher: an efficient client for managing multiple APs," in *Proc. of the second ACM SIGCOMM Workshop on Programmable Routers for Extensible Services of Tomorrow (PRESTO '09)*. ACM, 2009, pp. 43–48.

[135] S. Kandula, K. C.-J. Lin, T. Badirkhanli, and D. Katabi, "FatVAP: aggregating AP backhaul capacity to maximize throughput," in *Proc. of the fifth USENIX*

Symposium on Networked Systems Design and Implementation (NSDI '08), Apr. 2008.

[136] S.-L. Tsao and P. H. Lo, "DualMAC: A soft handoff mechanism for real-time communications in secured WLANs," *Computer Communications*, vol. 30, no. 8, pp. 1785–1793, Jun. 2007.

[137] U. Ramacher, "Software-defined radio prospects for multistandard mobile phones," *Computer*, vol. 40, no. 10, pp. 62–69, Oct. 2007.

[138] *Product Brief: PMB 2008, SMARTiTM WiMAX Single-chip dual-band WiMAX / WLAN RF Transceiver IC with standard I&Q interface*, Infineon Technologies AG. [Online]. Available: http://www.datasheetarchive.com/smarti*-datasheet.html

[139] *Product Brief: Intel® Centrino® Wireless-N + WiMAX 6150*, Intel. [Online]. Available: http://www.intel.com/Assets/PDF/prodbrief/324748.pdf

[140] L. Yang, V. Kone, X. Yang, Y. Liu, B. Zhao, and H. Zheng, "Coexistence-aware scheduling for wireless system-on-a-chip devices," in *Proc. of the 7th Annual IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks (SECON '10)*, 2010, pp. 1–9.

[141] H.-H. Choi, O. Song, Y.-K. Park, and J.-R. Lee, "Performance evaluation of opportunistic vertical handover considering on-off characteristics of VoIP traffic," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 3115–3121, Jul. 2010.

[142] M. Hollick, P. Mogre, C. Schott, and R. Steinmetz, "Slow and steady: Modelling and performance analysis of the network entry process in IEEE 802.16," in *Proc. of 15th IEEE International Workshop on Quality of Service (IWQoS '07)*, Jun. 2007, pp. 126–134.

[143] *IEEE 802.16-2009—Air Interface for Broadband Wireless Access Systems*, IEEE Std 802.16-2009, May 2009, revision of IEEE Std 802.16-2004.

[144] J. Zhu and H. Yin, "Enabling collocated coexistence in IEEE 802.16 networks via perceived concurrency," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 108–114, 2009.

[145] W. Yoon, "WiMAX and Bluetooth," *IEEE Vehicular Technology Magazine*, vol. 6, no. 4, pp. 60–67, 2011.

[146] M. Briggs, "Dynamic Frequency Selection (DFS) and the 5 GHz unlicensed bands," Whitepaper, Oct. 2010. [Online]. Available: http://www.elliottlabs.com/documents/dynamic_frequency_selection_combined.pdf

[147] *IEEE P802.11af—Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Amendment 5: TV White Spaces Operation*, IEEE Draft Std 802.11af/D6.0, Oct. 2013.

[148] R. Raghavendra, E. Belding, K. Papagiannaki, and K. Almeroth, "Unwanted link layer traffic in large IEEE 802.11 wireless networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 9, pp. 1212–1225, 2010.

[149] K. Yunoki and A. Ohara, "Real air-time occupation by Beacon and Probe," IEEE 802.11 TGai Fast Initial Link Set-up, Working Document 802.11-11/1413r3, Jan. 2012.

[150] ——, "Necessity of Probe reduction," IEEE 802.11 TGai Fast Initial Link Set-up, Working Document 802.11-12/0206r0, Feb. 2012.

[151] M. Anand, E. B. Nightingale, and J. Flinn, "Self-tuning wireless network power management," in *Proc. of the 9th ACM Annual International Conference on Mobile Computing and Networking (MobiCom '03)*, 2003, pp. 176–189.

[152] E. Rozner, V. Navda, R. Ramjee, and S. Rayanchu, "NAPman: network-assisted power management for WiFi devices," in *Proc. of the 8th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*, 2010, pp. 91–106.

[153] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* John Wiley & Sons Inc., May 1991.

[154] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 2, pp. 9–24, Apr. 2001.

[155] "OPNET Modeler." [Online]. Available: www.opnet.com

[156] Maxim Integrated, "Data Sheet: MAX2828/MAX2829 Single-/Dual-Band 802.11a/b/g World-Band Transceiver ICs," Oct. 2004. [Online]. Available: http://datasheets.maximintegrated.com/en/ds/MAX2828-MAX2829.pdf

[157] S. Wiethölter, "Virtual utilization and VoIP capacity of WLANs supporting a mix of data rates," TKN, TU Berlin, TKN Technical Report Series TKN-05-004, Sep. 2005.

[158] L. Cai, X. Shen, J. Mark, L. Cai, and Y. Xiao, "Voice capacity analysis of WLAN with unbalanced traffic," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 752–761, May 2006.

[159] J. G. Atallah and M. Ismail, "Future 4G front-ends enabling smooth vertical handovers," *IEEE Circuits and Devices Magazine*, vol. 22, no. 1, pp. 6–15, 2006.

[160] WiMAX Forum, "Mobile System Profile Specification, Release 1.5, Common Part," WiMAX Forum, Tech. Rep. WMF-T23-001-R015v01, Aug. 2009.

[161] A. Koepsel and A. Wolisz, "Voice transmission in an IEEE 802.11 WLAN based access network," in *Proc. of the 4th ACM International Workshop on Wireless Mobile Multimedia (WoWMoM '01)*, Jul. 2001, pp. 24–33.

[162] S. Shin and H. Schulzrinne, "Measurement and analysis of the VoIP capacity in IEEE 802.11 WLAN," *IEEE Transactions on Mobile Computing*, vol. 8, no. 9, pp. 1265–1279, Sep. 2009.

[163] A. Markopoulou, F. Tobagi, and M. Karam, "Loss and delay measurements of Internet backbones," *Computer Communications*, vol. 29, no. 10, pp. 1590–1604, Jun. 2006.

[164] *IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005—Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1*, IEEE Std 802.16e-2005 and IEEE Std. 802.16-2004/Cor1-2005, Feb. 2006, Amendment and Corrigendum to IEEE Std 802.16-2004.

[165] S. Wiethölter and A. Wolisz, "Selecting vertical handover candidates in WLAN hotspots," TKN, TU Berlin, TKN Technical Report Series TKN-08-009, Aug. 2008.

[166] ——, "Selecting vertical handover candidates in IEEE 802.11 mesh networks," in *Proc. of the first IEEE WoWMoM Workshop on Hot Topics in Mesh Networking (HotMESH)*, Jun. 2009, pp. 1–7.

[167] S. Wiethölter, M. Emmelmann, R. Andersson, and A. Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in *Proc. of the IEEE International Conference on Communications (ICC '12)*, Jun. 2012, pp. 5423–5428.

[168] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," *IEEE Network*, vol. 21, no. 1, pp. 27–33, 2007.

[169] C.-T. Chou, S. S. N, and K. Shin, "Achieving per-stream QoS with distributed airtime allocation and admission control in IEEE 802.11e wireless LANs," in *Proc. of the 24th IEEE International Conference on Computer Communications (INFOCOM '05)*, vol. 3, 2005, pp. 1584–1595.

[170] M. Davis and T. Raimondi, "A novel framework for radio resource management in IEEE 802.11 wireless LANs," in *Proc. of the third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt '05)*, 2005, pp. 139–147.

[171] C.-T. Chou, K. Shin, and N. Sai Shankar, "Contention-based airtime usage control in multirate IEEE 802.11 wireless LANs," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1179–1192, 2006.

[172] M. G. Patterson, "What is energy efficiency? : Concepts, indicators and methodological issues," *Energy Policy, Elsevier*, vol. 24, no. 5, pp. 377–390, 1996.

[173] M. Lacage, M. H. Manshaei, and T. Turletti, "IEEE 802.11 rate adaptation: a practical approach," in *Proc. of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '04)*, 2004, pp. 126–134.

[174] ITU-T Recommendation P.59, "Artificial conversational speech," Mar. 1993.

[175] Telecommunications Industry Association (TIA), "TSB-116-A: voice quality recommendations for IP telephony," Mar. 2006.

[176] V. Franz, "Ratios: A short guide to confidence limits and proper use," 2007, whitepaper.

[177] E. C. Fieller, "Some problems in interval estimation," *Journal of the Royal Statistical Society. Series B(Methodological)*, vol. 16, no. 2, pp. 175–185, 1954.

[178] U. von Luxburg and V. H. Franz, "A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations and bootstrap," *Statistica Sinica 19(3)*, pp. 1095–1117, 2009.

[179] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice capacity of IEEE 802.11b, 802.11a and 802.11g wireless LANs," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 3, 2004, pp. 1549–1553.

[180] F. Anjum, M. Elaoud, D. Famolari, A. Ghosh, R. Vaidyanathan, A. Dutta, P. Agrawal, T. Kodama, and Y. Katsube, "Voice performance in WLAN networks – an experimental study," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 6, 2003, pp. 3504–3508.

[181] E.-C. Park, D.-Y. Kim, and C.-H. Choi, "Analysis of unfairness between TCP uplink and downlink flows in Wi-Fi hot spots," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM '06)*, Dec. 2006, pp. 1–5.

[182] D. Dujovne, T. Turletti, and F. Filali, "A taxonomy of IEEE 802.11 wireless parameters and open source measurement tools," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, pp. 249–262, 2010.

[183] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama, "Voice capacity of IEEE 802.11b and 802.11a wireless LANs in the presence of channel errors and different user data rates," in *Proc. of the 60th IEEE Vehicular Technology Conference (VTC-Fall '04)*, vol. 6, Sep. 2004, pp. 4543–4547.

[184] S. Wiethölter, A. Ruttor, U. Bergemann, M. Opper, and A. M. Wolisz, "DARA: estimating the behavior of data rate adaptation algorithms in WLAN hotspots," in *Proc. of the 32nd IEEE International Conference on Computer Communications (INFOCOM '13), Mini Conference Track*, Apr. 2013, pp. 280–284.

[185] S. Wiethölter, A. Ruttor, U. Bergemann, M. Opper, and A. Wolisz, "DARA: estimating the behavior of data rate adaptation algorithms in WLAN hotspots," TKN, TU Berlin, TKN Technical Report Series TKN-13-001, Jan. 2013.

[186] The Madwifi Project Team, "The MadWifi Project." [Online]. Available: http://madwifi-project.org/

[187] W. Yin, K. Bialkowski, J. Indulska, and P. Hu, "Evaluations of MadWifi MAC layer rate control mechanisms," in *Proc. of IEEE International Workshop on Quality of Service (IWQoS '10)*, Jun. 2010.

[188] E. Ancillotti, R. Bruno, and M. Conti, "Experimentation and performance evaluation of rate adaptation algorithms in wireless mesh networks," in *Proc. of the 5th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks (PE-WASUN '08)*, 2008, pp. 7–14.

[189] I. Kim and Y.-T. Kim, "Realistic modeling of IEEE 802.11 WLAN considering rate adaptation and multi-rate retry," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1496–1504, Dec. 2011.

[190] M. Mirza, P. Barford, X. Zhu, S. Banerjee, and M. Blodgett, "Fingerprinting 802.11 rate adaption algorithms," in *Proc. of the 30th IEEE International Conference on Computer Communications (INFOCOM '11)*, Apr. 2011, pp. 1161–1169.

[191] P. A. K. Acharya, A. Sharma, E. Belding, K. C. Almeroth, and K. Papagiannaki, "Rate adaptation in congested wireless networks through real-time measurements," *IEEE Transactions on Mobile Computing*, vol. 9, no. 11, pp. 1535–1550, Nov. 2010.

[192] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[193] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[194] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Communications Magazine*, vol. 47, no. 3, pp. 88–95, Mar. 2009.

[195] Madwifi Project Team, *Minstrel*. [Online]. Available: http://madwifi-project.org/svn/madwifi/trunk/ath_rate/minstrel/minstrel.txt

[196] ——, *Snapshot archive for madwifi-hal-testing: MadWifi HAL testing, Release 4126*, Mar. 2010. [Online]. Available: http://snapshots.madwifi-project.org/madwifi-hal-testing/

[197] ——, *The Infamous Stuck Beacon Problems*. [Online]. Available: http://madwifi-project.org/wiki/StuckBeacon

[198] *IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems*, IEEE Std 1588-2008, Jul. 2008.

[199] "ptpv2d, v. 20101107." [Online]. Available: http://code.google.com/p/ptpv2d

[200] NetSpot Team, *Netspot 1.3: Free Wireless WiFi Site Survey Software for MAC OS X*. [Online]. Available: http://www.netspotapp.com

[201] "Iperf 2.0.4." [Online]. Available: http://sourceforge.net/projects/iperf/

[202] "FTP-Sever Pure-FTPd v. 1.0.28." [Online]. Available: http://www.pureftpd.org/project/pure-ftpd/

[203] "GNU Wget v. 1.12." [Online]. Available: http://www.gnu.org/software/wget/

[204] C. E. Rasmussen and H. Nickisch, "Gaussian Processes for Machine Learning (GPML) Toolbox," *Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, Dec. 2010.

[205] MATLAB, *version 7.13.0.564 (R2011b)*. MathWork Inc., Aug. 2011.

[206] T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimization*, vol. 6, pp. 418–445, 1996.

[207] ——, "On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds," *Mathematical Programming*, vol. 67, pp. 189–224, 1994.

[208] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[209] S. Wiethölter and C. Hoene, "An IEEE 802.11e EDCF and CFB simulation model for ns-2.26." [Online]. Available: http://www.tkn.tu-berlin.de/menue/softhardware_components/software/simulation_models

[210] ——, "Design and verification of an IEEE 802.11 EDCF simulation model in ns-2.26," TKN, TU Berlin, TKN Technical Report Series TKN-03-019, Nov. 2003.

[211] *The Network Simulator ns-2*. [Online]. Available: http://www.isi.edu/nsnam/ns/

[212] S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz, and L. Stibor, "IEEE 802.11e Wireless LAN for Quality of Service (invited paper)," in *Proc. of European Wireless 2002*, vol. 1, Feb. 2002, pp. 32–39.

[213] G. Ewing, K. Pawlikowski, and D. McNickle, "Akaroa 2.7.4 User's Manual," 2002.

[214] T. S. Rappaport, *Wireless Communications—Principles and Practice*. Prentice Hall, 2002.

[215] R. J. Punnoose, P. V. Nikitin, and D. D. Stancil, "Efficient simulation of Ricean fading within a packet simulator," in *Proc. of the 52nd IEEE Vehicular Technology Conference (VTC-Fall '00)*, vol. 2, Sep. 2000, pp. 764–767.

[216] E. Walker, H. Zepernick, and T.Wysocki, "Fading measurements at 2.4 GHz for the indoor radio propagation channel," in *Proc. of the International Zurich Seminar on Broadband Communications (IZS '98)*, Feb. 1998, pp. 171–176.

[217] A. Kochut, A. Vasan, A. U. Shankar, and A. Agrawala, "Sniffing out the correct physical layer capture model in 802.11b," in *Proc. of the 12th IEEE International Conference on Network Protocols (ICNP '04)*, 2004, pp. 252–261.

[218] M.-J. Ho, J. Wang, K. Shelby, and H. Haisch, "IEEE 802.11g OFDM WLAN throughput performance," in *Proc. of the 58th IEEE Vehicular Technology Conference (VTC-Fall '03)*, vol. 4, 2003, pp. 2252–2256.

[219] Cisco, "Cisco Wireless Mesh Access Points, Design and Deployment Guide, Release 7.3," Whitepaper, Aug. 2012.